

From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

DNA METHYLATION AND AGING: A LONGITUDINAL STUDY OF OLD SWEDISH TWINS

Yunzhang Wang



**Karolinska
Institutet**

Stockholm 2020

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet.

Printed by Eprint AB 2020

© Yunzhang Wang, 2020

ISBN 978-91-7831-634-2

DNA methylation and aging: a longitudinal study of old Swedish twins

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Yunzhang Wang

Principal Supervisor:

Dr. Sara H ägg
Karolinska Institutet
Department of Medical Epidemiology
and Biostatistics

Co-supervisor(s):

Dr. Åsa Katarina Hedman
Karolinska Institutet
Department of Medicine, Solna
Division of Rheumatology

Prof. Catarina Almqvist Malmros
Karolinska Institutet
Department of Medical Epidemiology
and Biostatistics

Dr. Malin Almgren
Karolinska Institutet
Department of Clinical Neuroscience
Center for Molecular Medicine

Dr. Robert Karlsson
Karolinska Institutet
Department of Medical Epidemiology
and Biostatistics

Opponent:

Dr. Bastiaan T. Heijmans
Leiden University Medical Centre
Department of Medical Statistics and
Bioinformatics

Examination Board:

Dr. Chengxuan Qiu
Karolinska Institutet
Department of Neurobiology, Care Sciences
and Society
Division of Aging Research Center

Dr. Pekka Katajisto
Karolinska Institutet
Department of Biosciences and Nutrition

Prof. Jan Dumanski
Uppsala University
Department of Immunology, Genetics and
Pathology

“逝者如斯夫，不舍昼夜”

"Tempus fugit"

ABSTRACT

DNA methylation is a well-known biomarker of aging. Many previous studies have reported the change of DNA methylation patterns with age, and analyzed DNA methylation in association with aging outcomes. However, most publications were based on cross-sectional data while longitudinal evidence was largely missing. Hence, in this thesis, we used longitudinal measures of DNA methylation from the Swedish Adoption/Twin Study of Aging (SATSA) to comprehensively study the role of DNA methylation in aging.

The first three studies in this thesis focus on different mechanisms of DNA methylation related to aging, including methylation level, methylation variability and epigenetic mutation.

In Study I, we investigated the longitudinal change of methylation level with age from an epigenome-wide association study (EWAS) using a mixed effect model. We identified 1316 age-related CpGs and successfully validated them in two external cohorts. Further, we analyzed the methylation difference between paired twins at the same time-point, and found it increased with age. We also identified genetic effect on age-associated CpGs, but the effect was independent on age.

In Study II, we first developed a method that could properly model the longitudinal change of methylation variability with age in simulated data. The method included a linear model to regress methylation on age, followed by a random intercept model to regress the absolute residuals on age. Next, we applied the method in an EWAS and identified 570 age-varying CpGs. The inter-individual variance of most CpGs increased with age longitudinally.

In Study III, we comprehensively studied epigenetic mutations, which are extreme outliers in the distribution of methylation level. The number of epigenetic mutations significantly increased with age in our longitudinal data. We also identified other factors associated with epigenetic mutations, including sex, B cell, sample quality, cancer diagnosis and first genetic principal component. Further, we classified CpGs into frequent mutated CpGs, highly methylated outliers (HMO) and lowly methylated outliers (LMO), and found frequent HMOs were more related to biological factors. In the end, we validated epigenetic mutations using bisulfite pyrosequencing and proved that epigenetic mutations were persist and could accumulate in aging.

In Study IV, we performed an EWAS to analyze methylation levels, methylation variability and epigenetic mutations in association with mortality. We observed age-varying effect of methylation level on all-cause mortality which may explain the poor replication in previous studies. We also identified CpGs of cancer genes related to death from cancer. In the end, we provided evidence that methylation variability could predict all-cause mortality.

LIST OF SCIENTIFIC PAPERS

- I. **Wang Y**, Karlsson R, Lampa E, Zhang Q, Hedman ÅK, Almgren M, et al. Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins. *Epigenetics*. 2018 Sep 2;13(9):975–87.
- II. **Wang Y**, Pedersen NL, Hägg S. Implementing a method for studying longitudinal DNA methylation variability in association with age. *Epigenetics*. 2018 Aug 3;13(8):866–74.
- III. **Wang Y**, Karlsson R, Jylh ä J, Hedman ÅK, Almqvist C, Karlsson IK, et al. Comprehensive longitudinal study of epigenetic mutations in aging. *Clin Epigenetics*. 2019 Dec 9;11(1):187.
- IV. **Wang Y**, Karlsson R, Karlsson IK, Hedman ÅK, Almgren M, Almqvist C, et al. DNA methylation in association to mortality: Evidence for time-varying and cause-specific effects. Manuscript.

OTHER RELATED PAPERS

- I. Sturm G, Cardenas A, Bind M-A, Horvath S, Wang S, Wang Y, et al. Human aging DNA methylation signatures are conserved but accelerated in cultured fibroblasts. *Epigenetics*. 2019 Jun 3;0(0):1–16.
- II. Wang Q, Wang Y, Lehto K, Pedersen NL, Williams DM, Hägg S. Genetically-predicted life-long lowering of low-density lipoprotein cholesterol is associated with decreased frailty: A Mendelian randomization study in UK biobank. *EBioMedicine*. 2019 Jul 1;45:487–94.
- III. Svane AM, Soerensen M, Lund J, Tan Q, Jylh ä J, Wang Y, et al. DNA Methylation and All-Cause Mortality in Middle-Aged and Elderly Danish Twins. *Genes*. 2018 Feb 8;9(2).
- IV. Karlsson IK, Ploner A, Wang Y, Gatz M, Pedersen NL, Hägg S. Apolipoprotein E DNA methylation and late-life disease. *Int J Epidemiol*. 2018 Jun 1;47(3):899–907.
- V. Karlsson Linn é R, Marioni RE, Rietveld CA, Simpkin AJ, Davies NM, Watanabe K, et al. An epigenome-wide association study meta-analysis of educational attainment. *Mol Psychiatry*. 2017 Dec;22(12):1680–90.

CONTENTS

1	Introduction	1
1.1	Epigenetics and DNA methylation	1
1.1.1	Epigenetics	1
1.1.2	DNA methylation	1
1.1.3	DNA methylation array.....	3
1.1.4	Epigenome-wide association study (EWAS)	4
1.2	DNA methylation in relation to aging	5
1.2.1	Molecular biology of aging.....	5
1.2.2	Age-related changes in DNA methylation	5
1.2.3	Epigenetic clock (DNA methylation age)	6
1.2.4	Methylation variability (Epigenetic drift)	7
1.2.5	Epigenetic mutation (Methylation outlier)	7
1.2.6	Mortality	8
2	Aims.....	9
3	Data sources.....	10
3.1	Study population.....	10
3.1.1	The Swedish Adoption/Twin Study of Aging (SATSA).....	10
3.1.2	External cohorts	10
3.2	Phenotypic data	11
3.3	DNA methylation data	12
3.3.1	Microarray data	12
3.3.2	Bisulfite pyrosequencing data.....	13
3.4	Genetic data	13
3.5	Gene expression and DNA methylation in cancers.....	13
4	Methods	15
4.1	Statistical methods.....	15
4.1.1	Modeling longitudinal data	15
4.1.2	Identification of cis-meQTLs	15
4.1.3	Data simulation and model evaluation	16
4.1.4	Identification of epigenetic mutations	16
4.1.5	Survival analysis	18
4.2	Regulatory features.....	18
4.3	Functional annotation	19
5	Results and Interpretation	20
5.1	Study I - DNA methylation levels change with age.....	20
5.1.1	Longitudinal EWAS on age.....	20
5.1.2	Genetic and age effect on DNA methylation	21
5.2	Study II - DNA methylation variability increases with age.....	23
5.2.1	Method development	23
5.2.2	Longitudinal EWAS of methylation variability and age	25
5.3	Study III – Epigenetic mutations increase with age	26

5.3.1	Epigenetic mutation in association with age and other factors.....	26
5.3.2	Validation in bisulfite pyrosequencing.....	27
5.4	Study IV – DNA methylation and mortality	29
5.4.1	EWAS on all-cause and cause-specific mortality	29
5.4.2	Methylation variability in association with all-cause mortality.....	30
6	General discussion and future perspective	32
7	Conclusion	34
8	Acknowledgement.....	35
9	Reference	37

LIST OF ABBREVIATIONS

5-mC	5-Methylcytosine
5-hmC	5-Hydroxymethylcytosine
bps	Base pairs
CpG	Cytosine phosphate Guanine
CVD	Cardiovascular disease
DAVID	Database for Annotation, Visualization and Integrated Discovery
DMR	Differentially methylated regions
DNAmAge	DNA methylation age
DNMT	DNA methyltransferase
DZ	Dizygotic
EWAS	Epigenome-wide association study
FDR	False discovery rate
GWAS	Genome-wide association study
HMO	Highly methylated outlier
IQR	Interquartile range
LBC	Lothian Birth Cohort
LMO	Lowly methylated outlier
MAF	Minor allele frequency
meQTL/mQTL	Methylation quantitative trait loci
MZ	Monozygotic
PC	Principal component
PIVUS	Prospective Investigation of the Vasculature in Uppsala Seniors
QC	Quality control
SATSA	Swedish Adoption/Twin Study of Aging
SNP	Single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
TET	Ten-eleven translocation
VMP	Variably methylated position

1 INTRODUCTION

1.1 EPIGENETICS AND DNA METHYLATION

1.1.1 Epigenetics

Epigenetics is a study of changes in gene functions that are heritable between cell generations without alterations in DNA sequence [1]. The term epigenetic was first introduced by Conrad Waddington [2], and is now widely used to describe all regulated processes from genetic material to the final product [3]. The definition of epigenetic covers a number of mechanisms in molecular biology. The widely-studied epigenetic mechanisms include DNA methylation, histone modification, non-coding RNA and so on. Both DNA methylation and histone modification organize chromatin structure, which determines the accessibility of the DNA sequence to transcription factors and therefore controls gene expression. The X chromosome inactivation in women is a typical example of an epigenetic regulation, where the additional X chromosome DNA is heavily packed in condensed heterochromatin for dosage compensation [4]. In addition, microRNAs are involved in RNA silencing and post-transcriptional regulation of gene expression [5]. All these epigenetic markers cooperate with each other and together present an epigenetic pattern that regulates gene expression and thus determines cell function. The distinct epigenetic patterns between different cell types create an epigenetic barrier which is the key to maintain cell identity. The epigenetic pattern is also inherited between cell generations so that the cell function is maintained after cell division.

Unlike DNA sequence, which does not change except somatic mutations, epigenetic markers can be modified due to various environmental exposures in human life time. How epigenetic markers change with different factors and further influence gene expression is the main research question in the field.

1.1.2 DNA methylation

DNA methylation is an epigenetic mechanism where methyl groups can be added to the DNA sequence with covalent bonds. DNA methylation mostly occurs on a cytosine followed by a guanine (CpG), where a methyl group is added to the 5' end of the cytosine to produce 5-methylcytosine (5-mC). There are other types of chemical modifications on cytosine, such as DNA hydroxymethylation (5-hmC), but methylation is the most common modification [6]. There are over 28 million CpG sites in the human genome and they are not evenly distributed. CpG-rich regions are named CpG islands which are about 300 to 3000 base pairs (bps) in length [7]. CpG islands have been found in about 70% of gene promoter regions located close to transcription start sites, and hence regulate gene expression [8]. However, about half of the CpG islands found in humans are located in gene bodies or intergenic regions [9].

The reaction of adding a methyl group to an unmethylated CpG can be catalyzed by three different DNA methyltransferases (DNMTs): DNMT1, DNMT3a and DNMT3b. DNMT1, which is also called the methylation maintainer, keeps DNA methylation pattern unchanged during cell replication by targeting hemi-methylated CpGs on the newly synthesized DNA

strand [10]. Therefore, DNA methylation can be inherited between cell generations allowing cells to maintain their cellular functions. The other two enzymes, DNMT3a and DNMT3b, are responsible for de novo DNA methylation [11], which is essential to establish DNA methylation pattern in embryonic development and cell differentiation [12,13]. On the other hand, the process of demethylation can be classified into passive and active processes [14]. The passive process refers to the dilution of 5mC in replication, where DNMT1 fails to maintain the methylation status so the cell-average methylation level decreases with replication rounds. The active demethylation process includes several steps of oxidation reactions catalyzed by ten-eleven translocation (TET) family enzymes, and followed by either passive replication dilution or base excision and repair processes (Figure 1).

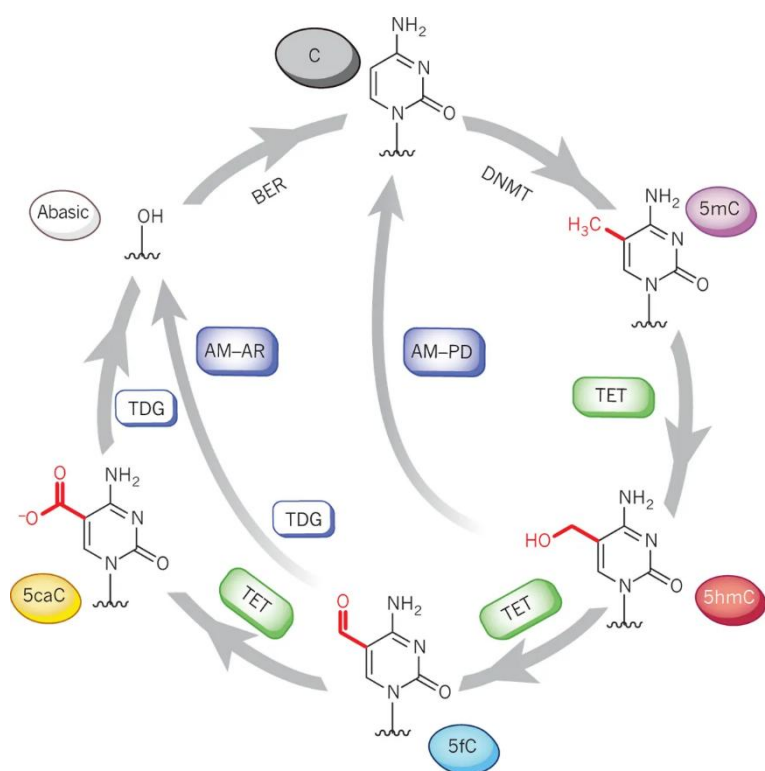


Figure 1. The pathway of DNA methylation and demethylation. The figure is modified from Kohli et al [14] with permission from Springer Nature.

DNA methylation is one of the most important epigenetic processes involved in regulating gene expression. The methylation status of different gene regulatory regions can regulate gene expression through different mechanisms. In promoter regions, DNA methylation can change the availability of gene promoters to transcription factors. In detail, methylated DNA can either inhibit the binding of transcription factors or recruit protein complexes that catalyze histone deacetylation, which results in a condensed chromatin and silencing of gene expression [15]. DNA methylation in enhancer regions can also reduce gene expression through similar mechanisms [16], and some evidences suggest enhancer methylation better correlates with gene expression than promoter methylation [17]. On the other hand, DNA methylation in gene body has bell-shaped correlation with gene expression, where both highly and lowly expressed genes have low levels of methylation in gene bodies [18].

DNA methylation patterns are different between cell types as they are part of the epigenetic barrier that preserve somatic cell identity [19]. Thus, it becomes an issue in association studies to distinguish true biological signals from what is caused by cellular heterogeneity [20]. As cellular compositions are not always measured in mixed cell samples, many algorithms have been developed to use highly cell-specific CpGs to predict cellular compositions. Typically, for whole blood samples, the Houseman method has been used to estimate compositions of leukocyte subtypes in combination with an external reference of purified leukocyte samples [21]. Alternatively, emerging reference-free methods have been developed to improve estimates of cellular compositions and remove the effect of cellular heterogeneity from methylation data [22].

The variance of DNA methylation levels across individuals is usually larger in CpGs of intermediate methylation levels, whereas CpGs of either high or low methylation levels show high consistency between and within individuals. The inter-individual variance of DNA methylation can be explained by genetic [23], environmental [24] and stochastic effects [25]. Growing evidence suggests that DNA methylation can mediate the genetic and environmental effects on complex diseases [26].

Studies on twins revealed that the monozygotic (MZ) twins have a higher correlation of DNA methylation than dizygotic (DZ) twins. The methylation heritability is estimated between 5% to 19% across the genome in different tissues [27–30]. CpGs with high heritability ($h^2 \geq 0.5$) usually have intermediate methylation levels and large variance across individuals [31]. To further analyze the genetic effect on methylation, studies on methylation quantitative trait loci (meQTL) have identified CpGs associated with single nucleotide polymorphisms (SNP) [23,32]. In general, the association becomes stronger when CpGs are located closer to their associated SNPs on the DNA sequence (cis-meQTLs). However, distal associations (trans-meQTLs) were also identified even between different chromosomes [33].

1.1.3 DNA methylation array

The Infinium Human Methylation 450K BeadChip, 450k array in short, is a microarray developed by Illumina, Inc. to perform a whole genome screening of DNA methylation levels in human. As suggested by the name, the 450k array is capable of measuring over 450 thousands CpGs. It is an extension of the predecessor 27k array. Recently, the 450k array started to be replaced by the EPIC array which can measure more than 850 thousands CpGs. One 450k array plate consists of 8 slides and each slide has 12 wells, so that in total 96 samples can be tested in a plate at a time.

To measure DNA methylation, a bisulfite reaction is first required to convert unmethylated cytosine to uracil, leaving methylated cytosine unchanged. Then, after polymerase chain reaction, unmethylated C-G pairs were turned into T-A pairs. Next, the 450k array can detect the C/T signals at specific CpG loci and therefore measure methylation levels. Specifically, there are two types of probes in the 450k array adopting different mechanisms to measure the signal. A type I probe has a pair of beads to detect methylated and unmethylated CpGs

respectively using the allele-specific oligonucleotides approach [34]. On the other hand, a type II probe only have one bead to detect both methylation status using the single-base extension approach [35]. Both types of probes produce fluorescent signals that represent methylated and unmethylated status. The two signals can be then calculated into beta-value, which is a variable between 0 and 1 representing the proportion of methylated CpGs in the study sample. The beta-value can be further transformed into M-value through a logit2-transformation. Beta-value is easier to interpret while M-value may perform better in statistical analysis [36].

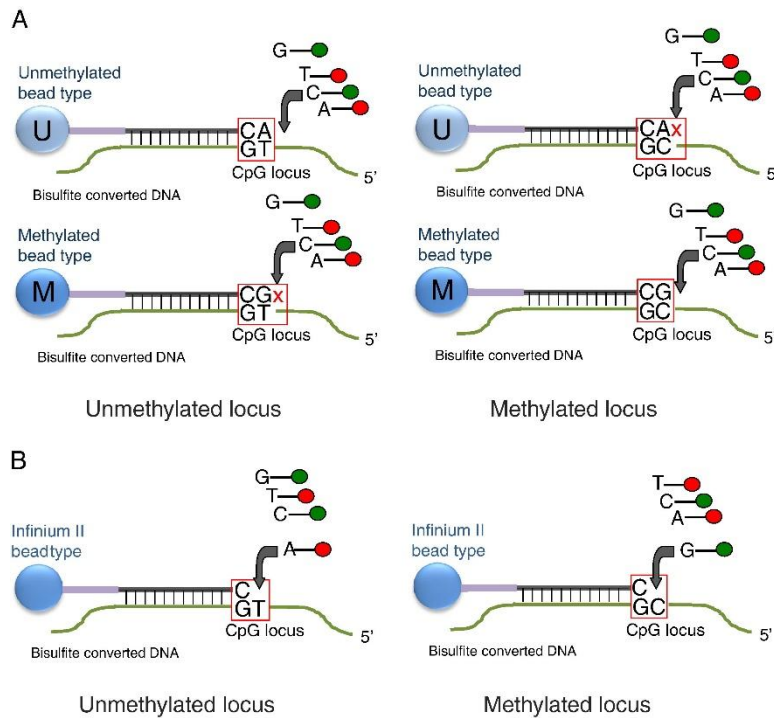


Figure 2. The mechanisms to measure DNA methylation in the 450k array including A) Type I probe and B) Type II probe. The figure is reprinted from Bibikova et al. [37] with permission from Elsevier.

The 450k array was designed to cover CpGs located in different regions of 99% Refseq genes [38]. However, since the human genome has over 28 million CpGs, only less than 2% of the whole methylome was measured by the 450k array. Thus, results of studies using 450k array are based on a small selection of human CpGs.

In terms of measure quality, Illumina claimed an over 98% reproducibility between technical replicates. But studies reported some probes of low intra-class correlation between replicates [39]. These CpGs mostly have low variance across samples, so they are more influenced by measurement errors. Also, low-variance CpGs are less likely to present biological signals from statistical analysis, so some studies simply removed them before the analysis.

1.1.4 Epigenome-wide association study (EWAS)

Epigenome-wide association study (EWAS) is a hypothesis-free design to exam the effect of DNA methylation on a complex trait, aiming to identify CpGs significantly associated with

the trait from the methylome. The idea of EWAS is similar to the genome-wide association study (GWAS), as they both applied the same statistical model on all measured SNPs or CpGs to test their associations with a trait [40].

Based on different measurement platforms, the total number of tests in an EWAS may vary from three to eight hundred thousand. Thus, the multiple testing problem needs to be considered in order to reduce false positive findings. In GWAS, a standard threshold was determined by the number of common independent variants in human genome [41]. However, there is no consensus on the number of independent tests in EWAS because of long-range correlations of DNA methylation between distal regulatory sites [42]. Most EWAS publications used the Bonferroni correction or the false discovery rate (FDR) [43] to determine a significance threshold. Also, as DNA methylation is more dynamic than DNA sequence, confounding environmental factors need to be adjusted in the EWAS model [40]. Moreover, repeated measures of DNA methylation can help estimate the individual trajectory over time, and can model the effect of exposures on the individual difference in the longitudinal change of DNA methylation.

1.2 DNA METHYLATION IN RELATION TO AGING

1.2.1 Molecular biology of aging

Aging is the process of losing physical and mental abilities with time, which ultimately leads to death [44]. It is the major risk factor for chronic diseases, such as cancer, dementia, diabetes, cardiovascular diseases (CVD) and autoimmune disorders [45]. To determine molecular bases of aging, studies have focused on lifespan-related gene mutations and functional decline in cell, tissue, organ and system levels [46]. At the cellular level, several mechanisms related to aging have been reported, including telomere shortening [47], declined genome maintenance [48], changes in epigenetic regulation [49], loss of protein homeostasis [50], deregulated nutrient sensing [51] and mitochondrial dysfunction [52]. These mechanisms together promote the time-dependent accumulation of cellular damage, which can lead to either cellular senescence [53] or uncontrolled cellular overgrowth [54]. The accumulation of senescent cells in aged tissue is associated with organismal aging [55], and unregulated cell division and increased cell heterogeneity are causes of cancer [56].

1.2.2 Age-related changes in DNA methylation

In the first three studies of this thesis, we studied the age-related alterations in DNA methylation including the change of methylation levels shared by individuals, which refers to age-related differentially methylated regions (aDMRs) [28]; the increasing divergence in methylation patterns with age, which is commonly called the “epigenetic drift” [25]; and the accumulation of methylation outliers, which is termed “epigenetic mutation” [57] (Figure 3).

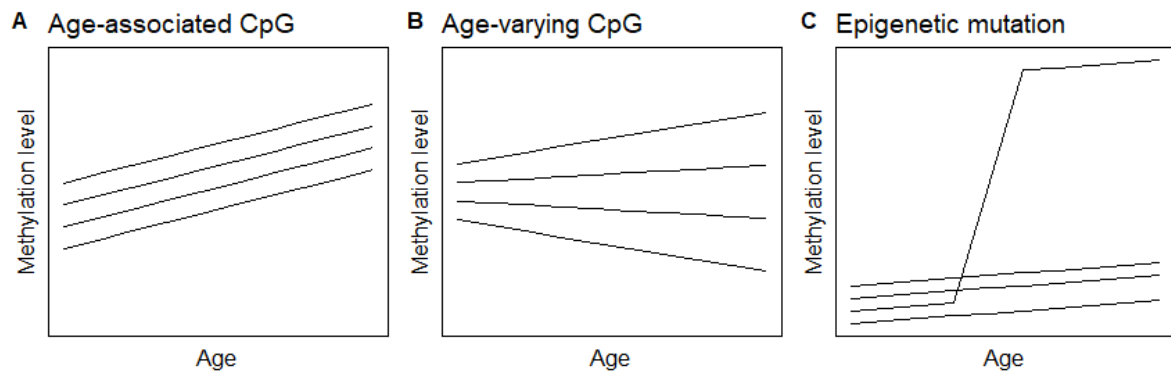


Figure 3. Schematic diagram shows the three types of alterations of DNA methylation related to aging. Each line represents the longitudinal change of DNA methylation in an individual. A) In age-associated CpGs, the average methylation levels change with age. B) Age-varying CpGs have methylation variability increase with age. C) Epigenetic mutations represent the development of methylation outliers with age.

Many studies have demonstrated that DNA methylation gradually change with chronological age throughout the human lifetime [58,59]. In general, the global DNA methylation level decreases with age (hypomethylation) [60], while the DNA methylation levels in CpG islands increase with age (hypermethylation) [61]. Both directions of age-related change of methylation are associated with the aging process. Age-related hypomethylation occurs in all genome regions [62], especially in regulatory protein binding sites [63], and contributes to genomic instability and cancer initiation [64]. In CpG islands, which usually harbor gene promoter regions and have low methylation levels, the age-related increase in methylation leads to inappropriate down-regulation of genes. For example, the promoter hypermethylations of tumor suppressor genes highly correlate with cancer risks [65].

Many publications of EWAS on age have identified large numbers of CpGs associated with chronological age using cross-sectional data [28,66–69]. Those results confirmed the global hypomethylations and specific hypermethylations in CpG islands. Overall, about 30% of CpGs measured by the 450k array showed strong or weak associations with age [31,68]. The strongest age-associated signal was found in the promoter region of the gene *ELOVL2*, where the DNA methylation level progressively increased during the whole lifespan [70]. However, the longitudinal evidence of individual change of methylation levels with age is very limited.

1.2.3 Epigenetic clock (DNA methylation age)

As a biomarker of aging, DNA methylation can be used to predict biological age, which is called epigenetic clock or DNA methylation age (DNAmAge). The predictor was usually built by first using a machine learning method to select informative CpGs related to age, and then using an a corresponding algorithm to incorporate them into an age estimate [71]. Among different types of biological age predictor, epigenetic clock based on DNA methylation is the most promising one in predicting health outcomes independent of chronological age [72].

The first two epigenetic clocks were developed by Horvath [73] and Hannum [66]. They both have high correlations with age, and the difference between the predicted DNAmAge and the chronological age can reflect the individual aging speed. Studies demonstrated that DNAmAge could predict all-cause mortality independent of other risk factors [74,75]. Also, increased DNAmAge was found associated with cancer incidence and mortality, but was less powerful to predict cardiovascular mortality, even though DNAmAge was predicted from blood samples [76]. Further age-related phenotypes and diseases were found to be associated with DNAmAge, include physical and cognitive fitness [77], frailty index [78], metabolic syndrome [79], Alzheimer's disease [80] and Parkinson's disease [81]. Till now, a few epigenetic clocks have been developed based on different tissues, algorithms and application purposes [82–84].

1.2.4 Methylation variability (Epigenetic drift)

Epigenetic drift describes the increasing divergence in DNA methylation between individuals. It can be caused by the deregulation of maintaining the normal methylation pattern and the accumulated effect of different environmental exposures over the human life course. Epigenetic drift is also a hallmark of aging [45] as it may explain the difference in individual aging rate [66].

Previous studies have observed epigenetic drift in aging from different perspective. Twin studies showed methylation differences between MZ twins increased with age [85,86]. Also, methylation variability, which was estimated by a test of heteroscedasticity, was reported to increase with age and age-related variably methylated positions (VMPs) were identified [66,87]. Moreover, epigenetic drift may plays an important role in aging related diseases, such as cancer [88,89]. However, there is a need to for longitudinal studies to follow the increase in methylation variability with age in the same people. Also, the study of methylation variability in relation to mortality is missing.

1.2.5 Epigenetic mutation (Methylation outlier)

Epigenetic mutation, which is known as aberrant methylation levels that could lead to unusual gene expression [90], is involved in cancer development and can also be an important factor in human aging. Unlike age-associated changes in methylation that are shared between individuals, epigenetic mutation is individual specific and the population average mostly stay unchanged. Epigenetic mutation also contributes to the increase in methylation divergence with age, but it is different to the methylation variability because epigenetic mutation focuses on rare outliers in random CpGs of conserved methylation levels. The accumulation of epigenetic mutations can lead to abnormal gene expression, genome instability, and cancer development in the aging process.

A previous study identified methylation outliers from samples and found that the total number of CpG sites with aberrant methylation levels increased exponentially with age [57]. A later study found significantly more epigenetic mutations in hepatocellular carcinoma tumors than in peritumoral samples, which suggests the importance of epigenetic mutations

in tumorigenesis [91]. However, longitudinal studies can better demonstrate if epigenetic mutations are persist over time and accumulate in the aging process.

1.2.6 Mortality

Being the final outcome of aging, mortality could also be predicted by DNA methylation. Emerging studies investigated DNA methylation in association with all-cause mortality, and built prediction models. The first EWAS reported 58 CpGs that could predict all-cause mortality independent of age, sex and smoking [52]. The identified CpGs were found in genes involved in various diseases including diabetes, CVD, various cancers and neuropsychiatric disorders. None of the mortality CpGs overlapped with previously identified age-related CpGs, but many of them were differentially methylated in smokers. Recently, more EWAS publications reported CpGs associated with all-cause mortality [92,93]. However, CpGs reported from these publications poorly replicated each other. Moreover, other than methylation level, methylation variability and epigenetic mutation may also predict mortality as they are biomarkers of aging.

2 AIMS

In Study I, we aimed to identify age-related CpGs from a longitudinal EWAS on age. We also aimed to investigate genetic and environmental effect on age-related CpGs from the analysis on twins and meQTLs.

In Study II, we aimed to develop a method that could properly model the longitudinal change of DNA methylation variability with age, and to apply it on longitudinal methylation data to identify age-varying CpGs.

In Study III, we aimed to comprehensively analyze epigenetic mutations in association with age and other factors.

In Study IV, we aimed to explore DNA methylation levels and methylation variability in association with all-cause and cause-specific mortality.

3 DATA SOURCES

3.1 STUDY POPULATION

3.1.1 The Swedish Adoption/Twin Study of Aging (SATSA)

The Swedish Adoption/Twin Study of Aging (SATSA) [94] is the main study cohort in this thesis. It is a cohort of old Swedish twin and aims to study genetic and environmental factors that cause individual differences in aging. SATSA was started in 1984 and the participants were followed every three years until 2014. Questionnaires and biological samples were collected from study participants longitudinally during the follow-ups. In this thesis, we measured DNA methylation data using blood samples from 402 participants, including 85 MZ and 116 DZ twin pairs. In total 1122 samples were measured at five time-points from 1992 to 2012. After quality control on methylation data, we retained 1011 samples from 385 participants for statistical analysis.

The characteristics of SATSA participants who contributed DNA methylation samples are shown in Table 1. SATSA participants were old with a mean age of 68.6. There were slightly more female participants than male. Some participants died during the follow-up time and new participants were recruited.

The genetic data of SATSA participants were also measured as described in section 3.4. Due to the QC on genetic data, in total 994 samples from 375 participants were available for both genetic and methylation measures. That resulted in a smaller sample size in Study III compared to the rest studies.

Table 1. The characteristics of SATSA participants whose DNA methylation data were used in this thesis.

Longitudinal wave	Year of sample collection	Number of Participants (new recruits)	Female Proportion	Age mean (SD)
1	1992-1994	239	59%	68.6 (9.1)
2	1999-2001	242 (102)	63%	71.2 (10.1)
3	2002-2004	188 (26)	54%	72.1 (9.1)
4	2008-2010	186 (15)	61%	76.2 (8.5)
5	2010-2012	156 (3)	66%	77.9 (8.4)

3.1.2 External cohorts

In Study I, two external cohorts were used to independently replicate results of longitudinal EWAS on age. In Both cohorts, longitudinal measures of DNA methylation from blood samples were available to exam the reproducibility of our results in a longitudinal perspective. The study characteristics of both external cohorts are shown in Table 2. Both cohorts measure DNA methylation from blood samples using 450k array.

The Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) [95] study included 196 Swedish people. In total 390 samples were collected from participants at age 70 and 80. Half participants in PIVUS were women.

The Lothian Birth Cohort (LBC) [96] study includes two birth cohorts, LBC1936 and LBC1921 [19]. A total of 1342 Scottish people were involved in the study. Participants of both sub-cohorts were measured up to three times. At the baseline, LBC1936 included 906 participants with a mean age of 69.6 years, and LBC1921 included 436 participants with a mean age of 79.1 years. The proportion of women was 49.4% in LBC1936 and 53.7% in LBC1921.

Table 2. Characteristics of the longitudinal DNA methylation samples from PIVUS and LBC.

Cohort	Longitudinal wave	Participants	Mean age	Female proportion
PIVUS	1	196	70.0	50%
	2	194	80.0	50%
LBC1936	1	906	69.6	49%
	2	801	72.5	48%
	3	619	76.3	48%
LBC1921	1	436	79.1	60%
	2	174	86.7	54%
	3	82	90.2	54%

3.2 PHENOTYPIC DATA

Phenotypic data were also collected from SATSA participants repeatedly during the follow-up. The phenotypic data used in this thesis included age, sex, twin zygosity and smoking status. The age, sex and zygosity information of the participants are described in section 3.1.1. The smoking status of participants were collected from questionnaires corresponding to the collection of blood sample. For the first available measure of the 385 participants, 304 were current-smokers, 15 were ex-smokers and 66 were never-smokers.

Some additional information was obtained by linking to registry data through personal ID, including date of cancer diagnosis, date of all-cause mortality and date of cause-specific mortality. The date of cancer diagnosis was retrieved from the National Patient Registry updated until May 2016, including ICD-codes for all cancer types (ICD7 codes 140-205, ICD8 codes 140-209, ICD9 codes 140-208, ICD10 codes C00-C97 and B21). In SATSA, 29 participants were diagnosed with cancer before the first observation, and 79 incident cases during the follow-up time.

The date of all-cause mortality were obtained from the Swedish National Register prior to May 2018. 236 participants died during the follow-up. The cause-specific mortality were obtained from the Cause of Death Registry prior to December 2016. A number of 46, 68, 30 and 46 participants died from cancer, CVD, stroke and dementia respectively.

3.3 DNA METHYLATION DATA

3.3.1 Microarray data

The 450k array was used to measure DNA methylation data used in this thesis (described in section 1.1.3). The raw methylation data obtained from methylation chip were pre-processed before statistical analysis. There are a number of approaches and software packages available for data preprocessing [97]. The preprocessing of the 450k array data used in this thesis included three steps: 1) quality control (QC), 2) normalization, 3) cell count and batch corrections. The preprocessing was implemented in the R package "RnBeads" [98].

The first step, QC, aims to remove low-quality samples and probes as they are not reliable. Usually, low-quality measurements have insufficient signals due to the lack of input DNA. The concept "detection p-value" created by Illumina can determine if a probe signal is sufficiently larger than the background noise measured by negative control probes. We adopted an algorithm named "greedy-cut" [98] to iteratively remove samples and probes with detection p-values over 0.05, optimizing the sensitivity and specificity. Another potential quality issue is sample contamination or mix-up. It can be detected by the 65 control probes that target SNPs instead of CpGs. We compared genotypes of 15 out of 65 SNPs available from a SNP array (described in section 3.4) with results from the QC probes, and calculated a correlation for each sample. Samples with a correlation lower than 0.7 were considered to be contaminated and were removed. Also, the ratio between probe signals from the X and Y chromosomes can be used to predict the sample sex, and we removed samples that were predicted to be the opposite sex. Moreover, probes that overlap with SNPs in their sequences were removed as they may not well hybridize to target DNA, resulting in unreliable measurements. In the end, we removed probes for sex chromosomes as we were not interested in them. After QC, we retained 390894 CpGs and 1011 samples for statistical analysis.

The second step of pre-processing was normalization. Raw methylation data need to be normalized to ensure that all the measurements follow the same distribution. A number of normalization methods are available to control color bias, subtract background signals and adjust difference between type I and type II probes [99,100]. In this study, we implemented the background correction method "noob" from the R package "methylnumi" [101] followed by a normalization method names "dasen" from the R package "watermelon" [102].

The third step included correction for cellular compositions and batch effect. The Houseman method [21] was used to predict cellular compositions. The method first identify top cell-specific CpGs from sorted blood-cell reference samples, and then use them to predict cellular compositions of study samples. We used 10 Swedish samples [103] as the reference panel and predicted the compositions of 9 blood cell types. Next, we used a simple linear regression to regress out the estimated cellular compositions from the normalized methylation data. Residuals from the model were transformed back to the scale of methylation data in beta-values.

Batch effect is a well-known type of bias caused by technical difference between experiments. Potential batch effect in methylation array can be the mean difference of measurements between plates, well positions, date of experiments and so on. We examined the distribution of our data and identified slides as the major batch effect. The "ComBat" method [104] implemented in the R package "sva" [105] was used to remove the known batch effect.

In the end, the methylation data were transformed into M-values for statistical analysis.

3.3.2 Bisulfite pyrosequencing data

Pyrosequencing is a method of DNA sequencing. Followed by the bisulfite conversion, pyrosequencing can also measure DNA methylation level as well. Compared to the 450k array, pyrosequencing only targets a specific region, usually a few hundred bps, but is capable of measuring the methylation levels of all CpGs within the region. In Study III, we used pyrosequencing as an alternative method to verify epigenetic mutations identified from the 450k array.

We selected 93 samples from 26 individuals to measure the methylation levels in four CpGs. The four CpGs are cg05270750, cg17338133, cg25351353, cg05124918. They were selected because they were frequently mutated in many samples and high-quality primers could be designed to cover them. Samples were then selected from participants of five measures to ensure at least five mutated samples in each of the four CpGs.

3.4 GENETIC DATA

The genetic data of SATSA participants were first directly genotyped using the Illumina PsychChip, which detected 588,454 SNPs for each individual. The QC on genetic data removed samples of high missing call rate, wrong relatedness between individuals and wrongly predicted sex, as well as SNPs of high missing call rate, not mapping to a chromosome and without minor allele.

After QC, the genetic data were imputed to predict unmeasured SNPs. Following a pre-phasing step, the genotyped data were imputed using IMPUTE2 with default parameters [106]. The imputation used the 1000 Genomes Project phase 1 [107] as the reference. Next, another QC on the imputed data removed SNPs of low quality ($\text{Info} < 0.6$) and low minor allele frequency ($\text{MAF} < 0.05$). Finally, the genetic data used in this thesis included 363 participants and over 6.5 million SNPs.

3.5 GENE EXPRESSION AND DNA METHYLATION IN CANCERS

In Study III, external RNA sequencing data of gene *PRDM7* were downloaded from The Cancer Genome Atlas Program (TCGA) through Wanderer [108]. The gene encodes a histone-lysine trimethyltransferase and was selected as it was related to a CpG (cg05270750) validated by pyrosequencing. We targeted the four most common types of cancer in human: lung cancer, breast cancer, colorectal cancer and prostate cancer. The specific cancer types available in TCGA were lung squamous cell carcinoma, lung adenocarcinoma, breast

invasive carcinoma, colon adenocarcinoma and prostate adenocarcinoma. For each cancer type, the gene expression were measured from both tumor and normal adjacent tissues. The sample size of all cancer types combined was 2209 for tumor tissue and 261 for normal adjacent tissue.

The corresponding DNA methylation data were also downloaded for these cancer types. There were 16 CpGs related to gene *PRDM7* available in TCGA, of which 15 CpGs were available in our methylation data.

4 METHODS

4.1 STATISTICAL METHODS

4.1.1 Modeling longitudinal data

Mixed effect model is a popular method to model longitudinal and related samples. In a mixed effect model, the effect of clustered groups, such as repeated measurements from the same individual, are adjusted as random effects by assuming samples from each group follow a normal distribution. Mixed effect model was widely applied in the first three studies of this thesis.

In Study I, we used a mixed effect model to regress DNA methylation levels on age. Sex was adjusted as a fixed effect. The longitudinal measures and twin pairs were adjusted as two random effects of random intercept, where the former is nested in the latter effect. It is also reasonable to assume that the rates of age-related change in methylation levels are different in people. However, adding a random slope in the model did not contribute much, but greatly increased the computational complexity in an EWAS study of half million regressions.

In Study II, we used a two-step approach including a linear model and a mixed effect model to measure the longitudinal change of inter-individual methylation variability, as an extension of Breusch–Pagan test [109] in longitudinal data. The method was tested in simulated data as described in section 4.1.3. In the first step, a linear model was used to regress methylation levels on age, and then a mixed effect model of random intercept was used to regress the absolute residuals from the first step on age. The simulated data proved that this method could properly measure the longitudinal change of variability with good statistical power.

In Study III, we used a mixed effect model to regress the longitudinal change of the number of epigenetic mutations with age. The outcome variable, the number of epigenetic mutations, was log-transformed so the distribution was close to normal. The fixed effects in the model included age, sex, B cell proportion, sample quality and the first genetic PC. Twin pairs and batch effect were adjusted as random effects. The repeated measures were not adjusted as it was nested in twin pairs and almost did not change the result. Although the batch effect was adjusted in QC, it was again adjusted as a random effect in the model because outlier analysis is sensitive to technical factors.

4.1.2 Identification of cis-meQTLs

The associations between genotypes and DNA methylation was studied and reported in Study I. In total 6.5 million imputed SNPs and CpGs were used in the test. To reduce the calculation complexity, we only identified cis-meQTLs, which means the associated SNP and CpG were located within 1 million bps. A linear model was used to regress methylation level on genotype adjusting for age, sex and the first four genetic PCs as covariates.

The identification of cis-meQTLs was implemented in two steps. First, the R package "MatrixEQTL" [110] was used to efficiently identify cis-SNP-CpG associations and calculate

a total of 1.94×10^9 tests. Next, the 2.5×10^6 associations with $p < 1 \times 10^{-5}$ were again tested in a linear model in combination with the robust standard error to adjust for sample correlation.

4.1.3 Data simulation and model evaluation

Simulated data were generated in Study II to evaluate the performance of methods for describing the longitudinal change of variability. The data simulated the longitudinal change of methylation levels with age in 30 subjects. The baseline ages of subjects were generated from a normal distribution. Subjects were measured in five waves for every 5 years. The baseline methylation levels were generated from a normal distribution. The individual rate of the change in methylation level with age was positively correlated with the intercept. In short, the simulation was based on a random intercept and slope model, where the intercept and slope were correlated. Therefore, the variability of methylation levels increased with age at a constant rate. In order to statistically compare model behavior, data were simulated 100 times with the same parameters but different random number generators for each test.

After selecting the optimal method, we changed the parameters of data simulation to test the model behavior under different circumstances, including the length of follow-up intervals, numbers of follow-ups and the number of subjects. The distributions of estimates and t-statistics from the method showed how these factors influenced the model results.

4.1.4 Identification of epigenetic mutations

In Study III, the term epigenetic mutations were used to describe outlier samples in the distribution of DNA methylation. The definition of being an outlier followed a previous publication [57], which is three times interquartile range (IQR) lower than the first quartile or higher than the third quartile (Figure 4). For each CpG, we identified outliers as epigenetic mutations from the distribution of beta-values. Then we counted the total number of epigenetic mutations in samples, which was the primary outcome variable in the study.

Age and sex are the two major factors in DNA methylation. The identification of epigenetic mutations was stratified by sex to avoid the misclassification of outliers due to sex difference. On the other hand, from the longitudinal data, only the first available measure of each individual were used to identify epigenetic mutations. There were two reasons to use the first observation without stratification by age: 1) It provided a distribution of independent samples with a maximized sample size; 2) The samples of the first observation were relatively young, so they provided a methylation distribution of relatively normal and healthy samples. This was consistent with the study hypothesis that epigenetic mutations were developed and accumulated in the aging process.

As the methylation level of some CpGs are highly influenced by genetics, the outliers in these CpGs may simply have a different genotype, rather than being caused by biological factor related to aging. Thus, we excluded CpGs related to cis-meQTL (described in section 4.1.2) from counting the number of epigenetic mutations in samples.

Methylation outliers can have two directions in different CpGs, either higher or lower than average methylation levels. The term highly methylated outlier (HMO) was used to describe outlier samples that had methylation levels much higher than the majority, and the term lowly methylated outlier (LMO) for the opposite cases (Figure 4).

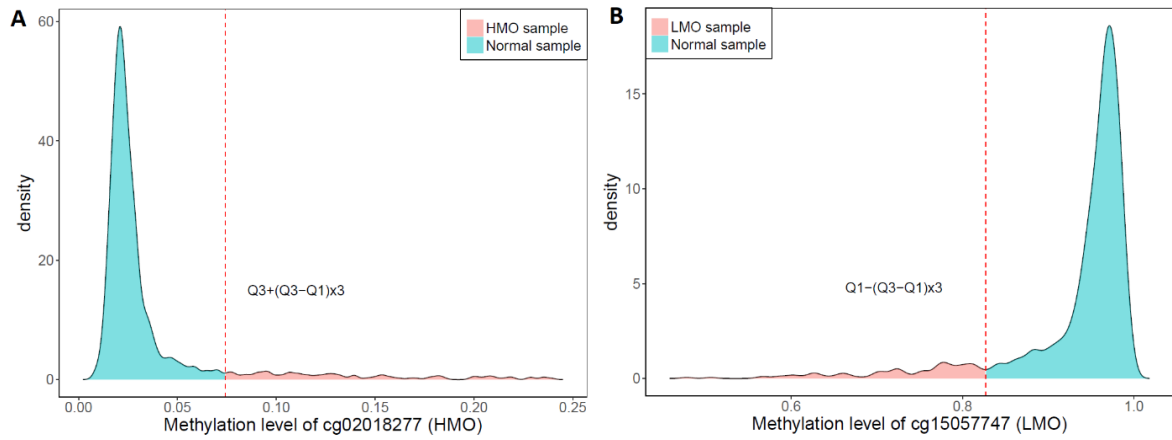


Figure 4. The distribution of A) cg02018277 and B) cg15057747 are examples of epigenetic mutations in two directions.

The number of outlier samples in CpGs has a skewed distribution, where most CpGs have a small number of outliers. However, some CpGs were more vulnerable to epigenetic mutations and thus their properties were interesting (Figure 5). To specifically study them, CpGs of more than 50 outlier samples were called frequently mutated CpGs. Then, the number of frequently mutated CpGs, and in combination with HMO/LMO, were calculated for each sample.

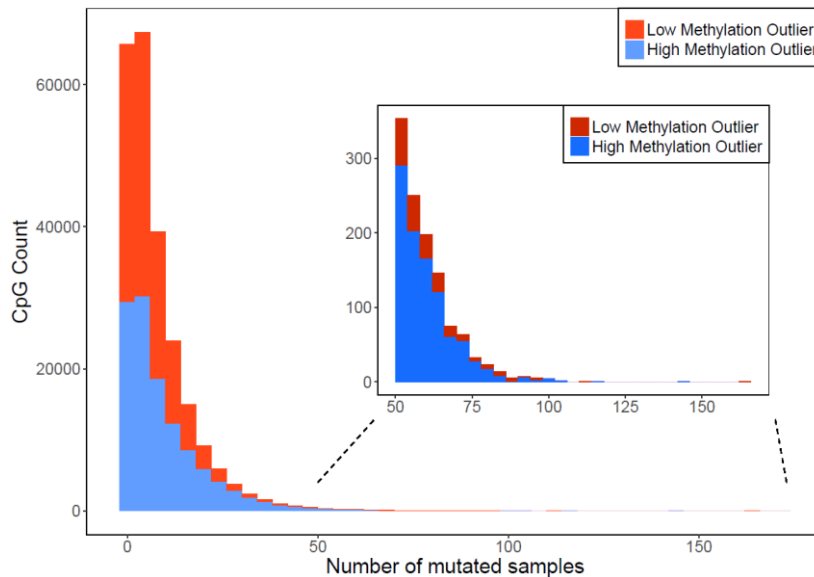


Figure 5. The distribution of the number of outlier samples in CpGs, stratified by high methylation outliers (HMOs) and low methylation outliers (LMOs).

4.1.5 Survival analysis

In Study III, the Cox proportional hazard (Cox) model was used to estimate the effect of epigenetic mutations on cancer diagnosis. The longitudinal number of epigenetic mutations was used as a time-varying covariate. Attained age was adjusted as the underlying time scale. Sex and baseline smoking status were adjusted as covariates. Both twin pair and batch effect were adjusted by robust standard error.

In Study IV, the Cox model was used to analyze methylation levels and methylation variability in association with all-cause and cause-specific mortality respectively. Attained age was adjusted as the time scale. Sex and smoking status were adjusted as covariates. Twin pairs were adjusted by robust standard error. To apply the Cox model on longitudinal measures, DNA methylation was modeled as a time-varying covariate. For each event, the most recent measurements of all available participants contributed to the prediction. An alternative way of modeling longitudinal exposure on time-to-event outcome is the joint model, which is a combination of a mixed effect model that models the longitudinal exposure, and a survival model that uses estimated exposure from the mixed model to predict the outcome. Theoretically, the joint model is the better choice to model the longitudinal methylation effect on mortality, as it makes use of longitudinal information of DNA methylation. However, in practice, the joint model was much more complicated and slower than the Cox model, making it inappropriate for an EWAS of half-million tests.

In Study IV, we also tested the time-varying effect of methylation on all-cause mortality. We added a covariate of age-methylation interaction to model a linear change of methylation effect over age. The age in the interaction covariate was centered at age 70, so that the estimate from the methylation represented the effect at age 70.

4.2 REGULATORY FEATURES

To interpret biological functions of significant CpGs identified from EWAS, we annotated regulatory features of target CpGs to explore their roles in gene regulation. In this thesis, the genome annotation of regulatory features was sourced from the Ensembl Regulatory Build [111]. Genomic regions involved in gene regulation are classified into Promoters, Promoter flanking regions, Enhancers, CTCF binding sites, Transcription factor binding sites and Open chromatin regions. The classification was based on experimental data from open chromatin assays, histone modification assays and transcription factor binding assays. These experimental data provided patterns of chromosome availability and epigenetic modifications, which were used to predict cell-type specific chromatin states. States from all available cell types were then combined to predict regulatory features [112] (Figure 6).

In practice, to annotate the regulatory feature of a CpG, we created a 50bps slice that covered the target CpG in the center and searched for regulatory regions overlapping with the slice. To ensure reproducibility, we used the regulatory build released in November 11th, 2016.

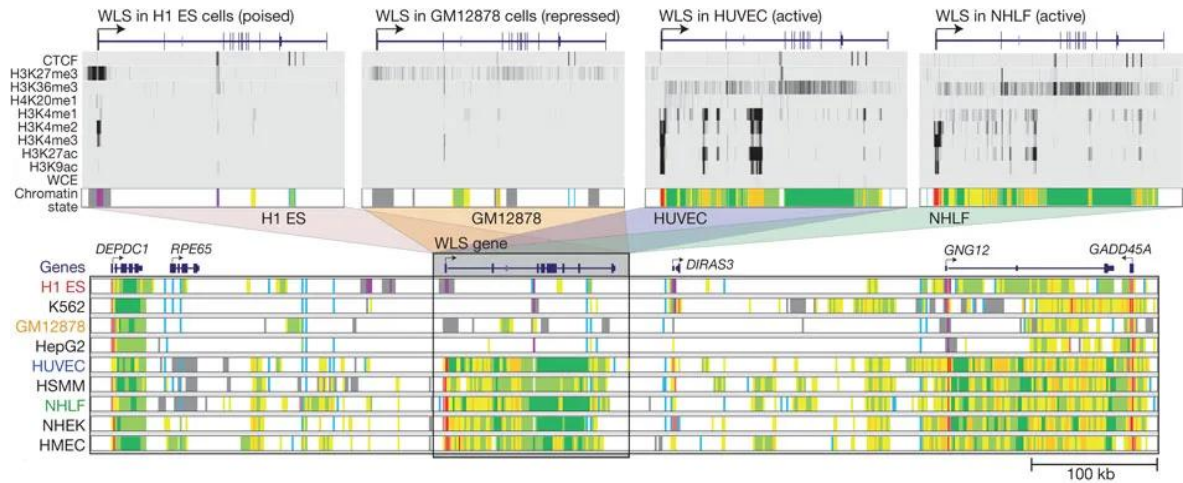


Figure 6. The process of predicting regulatory features from experimental data. The figure is modified from Ernst et al. [112] with permission from Springer Nature.

4.3 FUNCTIONAL ANNOTATION

To explore the function identified CpGs and their related genes, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID) online tool [113,114] to annotate the gene ontology terms in this thesis. The related gene of a target CpG is available from the 450k manifest file. The DAVID tool can take a list of genes with a background reference to look for enrichment in cellular functions, pathways and disease associations. The enrichment was calculated based on the co-occurrence with genes in functional categories. The gene ontology terms was reported in this thesis to describe the attribute and potential functions of the interesting genes.

5 RESULTS AND INTERPRETATION

5.1 STUDY I - DNA METHYLATION LEVELS CHANGE WITH AGE

5.1.1 Longitudinal EWAS on age

From the longitudinal EWAS on age, we identified 1316 CpGs of which the methylation levels were significantly associated with age under Bonferroni correction ($p < 1.3 \times 10^{-7}$). The methylation level of the top finding cg16867657, which locates in the promoter region of the gene *ELOVL2*, longitudinally increased with age ($p = 2.3 \times 10^{-31}$). The strong age association of cg16867657 has been reported before [70] and was later verified in vitro [115].

Our results were overall consistent with the previous cross-sectional studies using 450k data (Table 3). The number of significant results from studies varied largely from hundreds to ten thousands, due to different study designs and choices of threshold. Our finding were mostly covered by studies which reported a large number significant CpGs, especially in Johansson, 2013 [68] that also used Swedish data.

The age-associated CpGs were further validated in two external cohorts with longitudinal measures of DNA methylation (described in section 3.1.2). Under a Bonferroni-correction threshold, more than half of our results were verified in LBC but only less than 10% were verified in PIVUS. The correlation of effect sizes was 0.87 between SATSA and LBC, and 0.57 between SATSA and PIVUS. The difference in replication performance was possibly because study characteristics of SATSA was more similar to LBC than PIVUS. Compared to PIVUS, LBC had a larger sample size, wider age range, longer follow-up time and more replicated measures.

Table 3. Summary of age-related CpGs identified in SATSA and in previous publications.

Study	Significant CpGs	Analyzed in PIVUS/LBC	Validated in PIVUS/LBC	Study type
SATSA	1316	-	-	Longitudinal
PIVUS	-	1271	118	Longitudinal
LBC	-	973	594	Longitudinal
		Analyzed in SATSA	Validated in SATSA	
Florath, 2013 [69]	162	153	81	Cross-sectional
Johansson, 2013 [68]	137,993	116,042	1,192	Cross-sectional
Dongen, 2016 [31]	135,775	117,406	873	Cross-sectional
Tan, 2016 [116]	2,284	2,087	302	Longitudinal
Horvath, 2013 [73]	353	316	15	Epigenetic-clock
Hannum, 2013 [66]	71	63	33	Epigenetic-clock

Age-associated CpGs identified in this study were more located in CpG shores and less in CpG islands and open sea regions, compared to all CpGs after QC. Most hypermethylated CpGs (85.2%) were located in CpG islands (Figure 7A). The annotation of regulatory

features showed that age-associated CpGs were enriched in CTCF binding sites, promoter flanking regions and other transcription factor binding sites (Figure 7B). Different to previous beliefs, age-related hypermethylated CpGs were not strongly enriched in promoter regions. Instead, they were more located in TF binding sites without known regulatory features. The enhancers and open chromatin regions were poorly covered by the 450k array.

We further mapped the age-associated CpGs to 878 genes and annotated their gene functions. Seven enriched GO terms were identified under $FDR < 0.05$. The top functions included homophilic cell adhesion via plasma membrane adhesion molecules, nervous system development and neurogenesis. Thus, genes regulated by age-associated CpGs could have important functions in inflammatory response and nervous system.

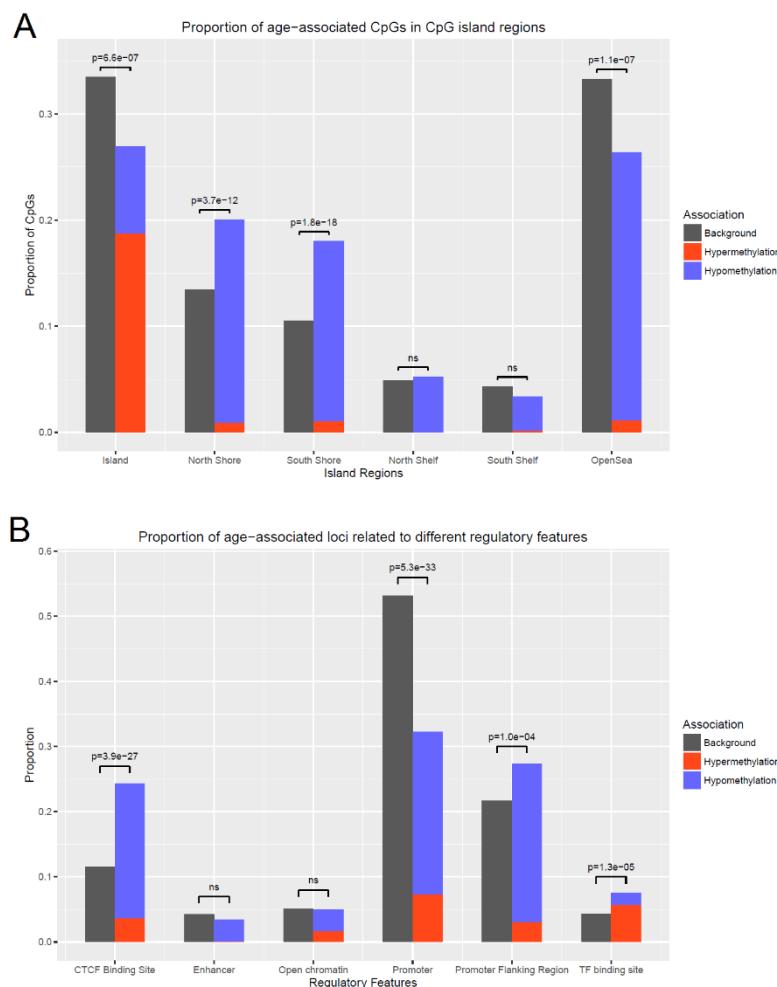


Figure 7. The distributions of age-associated and background CpGs in A) CpG islands and B) regulatory features.

5.1.2 Genetic and age effect on DNA methylation

Our study on cis-meQTLs reported over 1.4 million significant associations under Bonferroni correction ($p < 2.5 \times 10^{-11}$). Stronger associations were observed when the associated SNP and CpG were closer each other (Figure 8). In total 14,714 CpGs were significantly associated with at least one SNP, suggesting around 3.7% CpGs from 450k array are highly influenced

by genetic. Our results were in general consistent with mQTL database [32], a previously published study on meQTLs. About 44% of our identified SNP-CpG associations were validated ($p < 10^{-14}$) in the middle-age group from the mQTL database.

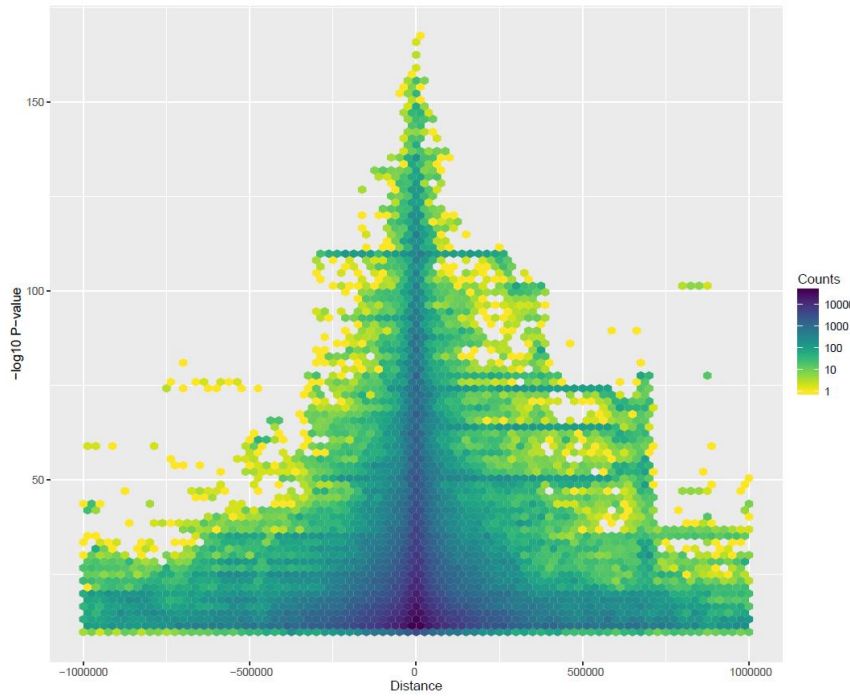


Figure 8. The p-values of associations between CpGs and cis-SNPs in relation to their distances in the genome. More associations (blue) and lower p-values were observed when associated SNP-CpG pairs are closer.

We further studied how age and genetic effects on DNA methylation interplay. Of the 1316 age-associated CpGs, 123 (9.3%) were associated with at least one SNP, which was higher than the proportion of SNP-associated CpGs in all CpGs (3.7%; $p = 6.7 \times 10^{-26}$). We further included related SNPs in the GWAS model of these CpGs. The age effects mostly stayed unchanged and no interaction was observed between age and SNPs.

We also used the twin design to study genetic and age effects. We calculated standardized Euclidean distances of DNA methylation between paired twins measured at the same time, and regress the distance on age, sex and zygosity. Taking all CpGs into account, the methylation difference between twins increased significantly with age ($\beta = 0.021$, $p = 9.4 \times 10^{-4}$, Figure 9A). The age effect was stronger by calculating the distance of 1316 age-associated CpGs ($\beta = 0.029$, $p = 2.9 \times 10^{-5}$; Figure 9B). For SNP-associated CpGs, the age effect was smaller but still statistical significant ($\beta = 0.015$, $p = 3.32 \times 10^{-5}$; Figure 9C). Therefore, age-associated CpGs could have higher age-induced variations as well, implying that they are more influenced by environmental and stochastic effects.

On average, the methylation difference between DZ paired twins was 5.2% higher than that of MZ paired twins using all CpGs. This number was 9.8% for age-associated CpGs and 42.8% for SNP-associated CpGs. We did not detect an age-zygosity interaction. Overall, results from

the twin analysis agreed with results from meQTL, as both suggested stronger genetic effect on age-related CpGs than average.

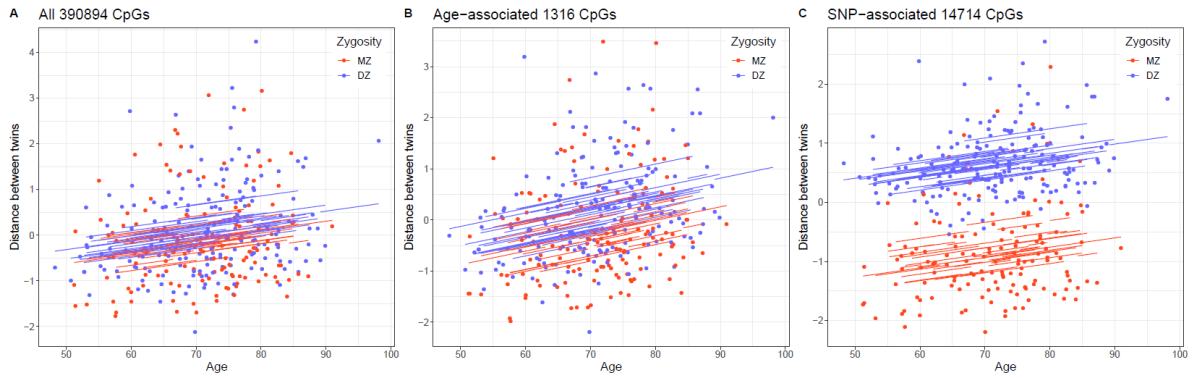


Figure 9. The methylation differences between paired twins in association with age and zygosity in SATSA. Methylation differences were calculated from Euclidean distances of A) all CpGs, B) age-associated CpGs, and C) SNP-associated CpGs.

5.2 STUDY II - DNA METHYLATION VARIABILITY INCREASES WITH AGE

5.2.1 Method development

We performed a simulation study to determine an appropriate method based on the Breusch–Pagan test [109] to estimate the inter-individual methylation variability in longitudinal data. The generation of the simulated data is described in section 4.1.3. The Breusch–Pagan test includes two steps of regression: first regress out the independent variable, and then model the absolute residuals from the first step on the independent variable to estimate the change of variance.

For the first step, we tested a simple linear model (Model 1.1, Figure 10A), a random intercept and slope model (Model 1.2, Figure 10E), and a random intercept model (Model 1.3, Figure 10G). Only residuals from Model 1.1 captured inter-individual variability as the absolute residual increased with age. In the second step, we tested the same three models to regress the absolute residuals on age (Figure 10B-D). Model 2.3 was chosen as it had the best statistical power which is important in EWAS. Therefore, we concluded that a linear regression followed by a random intercept model could properly measure the longitudinal change of inter-individual variability over time.

We also tested the model performance under different circumstances. We simulated datasets of different age ranges and follow-up intervals. The estimated age effect on variability stayed unchanged, but the longer follow-up time provided better statistical power. We also tested the model with different numbers of participants and repeated measures. Again, the estimated age effect stayed unchanged, but both the number of participants and the number of measures were positively associated with the test power.

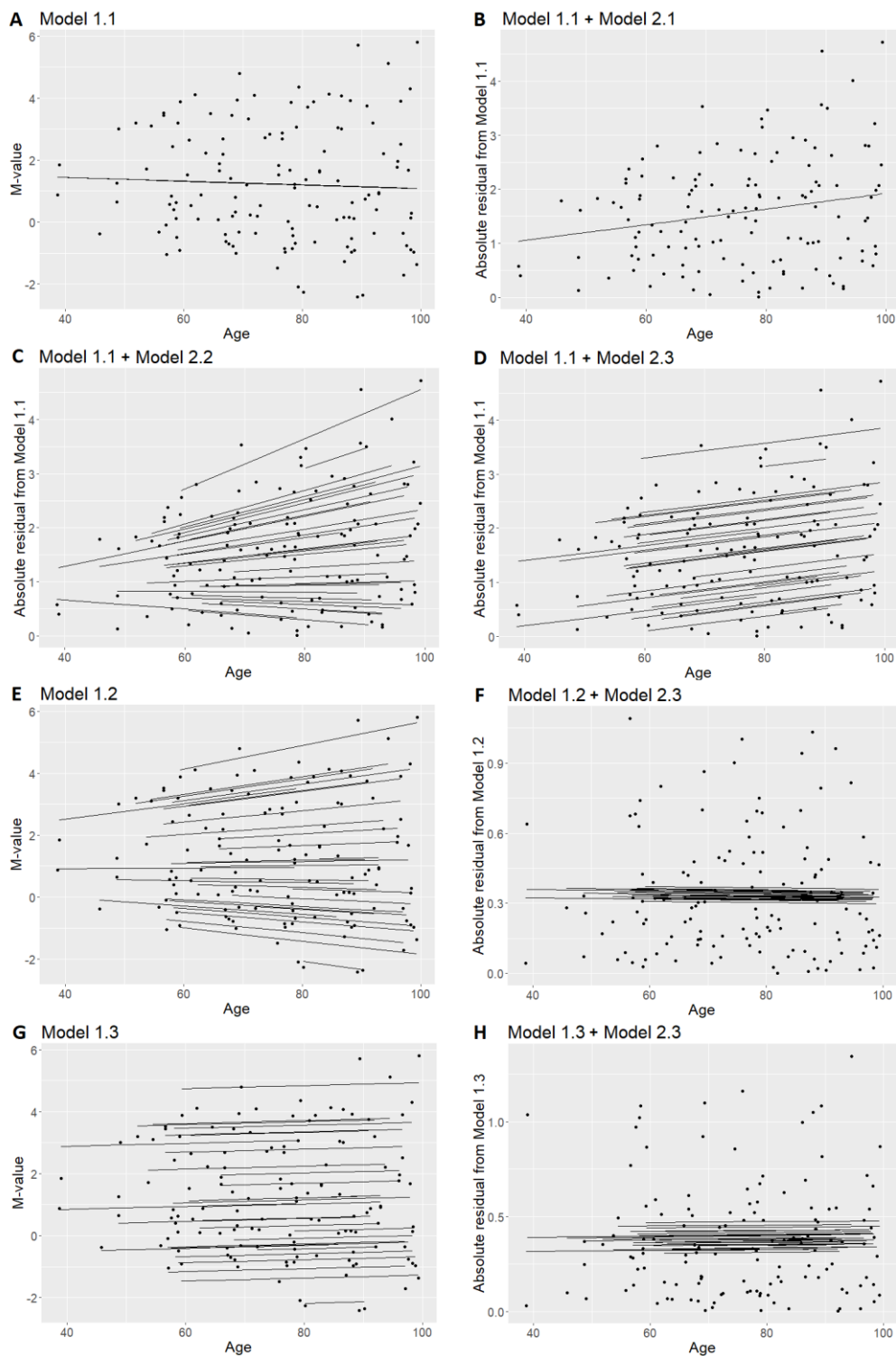


Figure 10. Models tested on simulated data to determine an appropriate method for testing heteroscedasticity in longitudinal data. A) The simple linear model (Model 1.1) generated absolute residuals that measured inter-individual variability. The absolute residuals were then regressed on age by B) a simple linear regression (Model 2.1), C) a random intercept and slope model (Model 2.2) and D) a random intercept model (Model 2.3). E) The random intercept and slope model (Model 1.2) best fitted the simulated data. F) But the absolute residuals captured intra-individual variability. G) The random intercept model (Model 1.3) was also tested. H) But absolute residuals were not associated with the age.

5.2.2 Longitudinal EWAS of methylation variability and age

After determining the proper method, we applied the method in an EWAS to estimate the change of inter-individual variability of DNA methylation with age. The results showed that 90.4% of all CpGs had positive effect sizes, indicating the variability of most CpGs increased with age. We identified 570 significant age-varying CpGs after Bonferroni correction ($p < 1.3 \times 10^{-7}$). These age-varying CpGs were mapped to 246 genes and the functional analysis of those genes in DAVID online tool [113] showed them enriched in the GO term “nervous system development” ($p = 1.9 \times 10^{-5}$, FDR=0.034). Our study results were consistent with a previous cross-sectional study [87], which replicated 218 of the 570 CpGs identified in our study. Their functional analysis also reported enriched gene function in neuron development. Considering the smaller sample size (1011 samples from 385 individuals versus 3295 samples) and narrower age range (48 to 99 year versus 10 to 90 year) in our study, the high replication rate suggests that our method was correct and robust.

We further compared them with the age-associated CpGs reported in Study I, and found 7 CpGs to be both age-associated and age-varying (Figure 11A and B). We also found 14 age-varying CpGs associated with genetic variants, including one (cg06464078) of which the methylation variability decreased with age (Figure 11D). Basically, the average methylation levels of most age-varying CpGs stayed unchanged with age, which is similar to the previous study [87]. This implies that age-related and age-varying CpGs may have different mechanisms in aging.

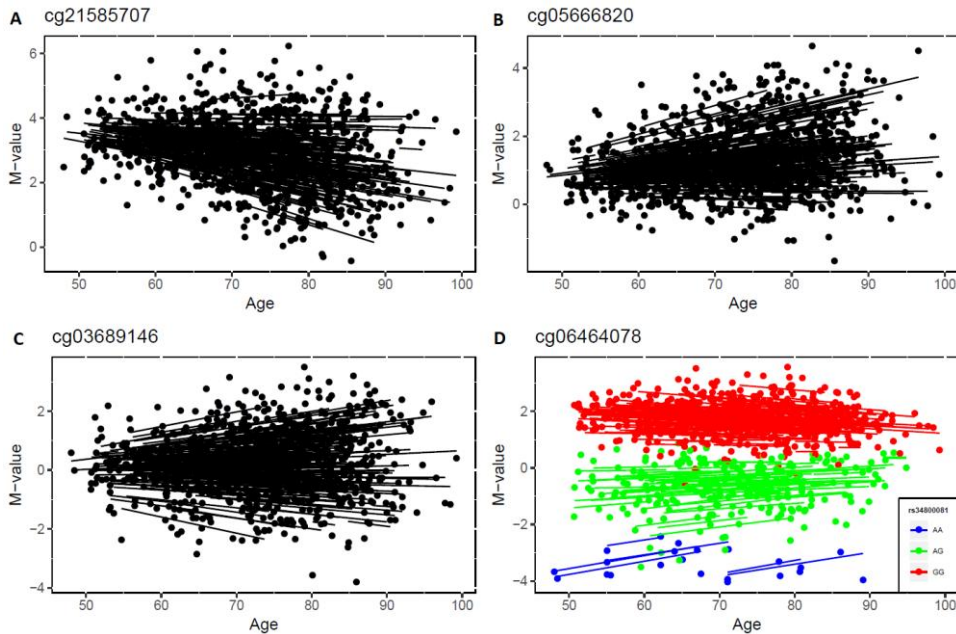


Figure 11. Examples of longitudinal change of DNA variability with age. A) The methylation levels of cg21585707 decreased with age but the variability increased with age. B) Both methylation level and methylation variability of cg05666820 increased with age. C) The methylation variability of cg03689146 increased with age but the average methylation level stayed unchanged. D) The methylation variability of cg06464078 decreased with age, and was also associated with a SNP rs34800081.

5.3 STUDY III – EPIGENETIC MUTATIONS INCREASE WITH AGE

5.3.1 Epigenetic mutation in association with age and other factors

Epigenetic mutations were identified and counted in SATSA as described in section 4.1.4. We verified the cross-sectional finding that the number of epigenetic mutations exponentially increased with age ($p=1.22 \times 10^{-13}$) in longitudinally data. Additionally, we found other factors associated with the number of epigenetic mutations (Table 4), including female ($p=6.33 \times 10^{-3}$), low sample quality ($p=1.48 \times 10^{-117}$), CD19+ B cell composition ($p=5.06 \times 10^{-23}$), cancer diagnosis ($p=0.014$) and the first genetic PC ($p=0.041$).

As epigenetic mutations are outliers, they were expected to be strongly affected by technical factors including sample quality and batch effect. Although low-quality measurements were largely removed in QC, sample quality still plays a major role in the number of epigenetic mutations. The same applies to the batch effect so we additionally adjusted batch as a random effect. As a result of sample randomization on the 450k array, technical factors did not confound the age effect on epigenetic mutations in this study. However, for a single CpG, it is difficult to distinguish outliers induced by technical factors from that caused by biological factors. The strong effect of CD19+ B cell on epigenetic mutations could be explain by the unique methylation pattern of B cell compared to other lymphocytes [103]. The minor genetic factor could be caused by CpGs associated with trans-meQTLs which were not removed.

We further classified them into HMOs, LMOs and frequently mutated CpGs as described in section 4.1.4. We have found 1,185 (0.32%) frequently mutated CpGs, and two of them were associated with age identified in Study I. The number of frequently mutated CpGs were also significantly associated with age, sample quality, CD19+ B cell compositions, cancer diagnosis and genetic PC1, but not with sex (Table 4).

More HMOs were found in frequently mutated CpGs than LMOs (969 and 216, $p < 1 \times 10^{-16}$, Figure 5), which is consistent with a previous publication [117]. We further studied frequent HMOs and frequent LMOs, and found stronger effects of age, B cell, genetic PC1 on the former (Table 4). However, frequent LMOs were much more affected by sample quality than frequent HMOs (Table 4).

The different effect of covariates on frequent HMOs and LMOs suggested that the direction of epigenetic mutation was important. Frequent HMOs were more associated with biological factor like age and B cells. On the other hand, the number of frequent LMOs was small but more affected by sample quality, making it difficult to study their biological functions.

Moreover, a survival analysis of epigenetic mutations in association with cancer diagnosis showed that a higher number of frequent HMOs could be a risk factor ($p=0.048$). However, the total epigenetic mutations and frequent LMOs were not associated with cancer diagnosis.

Table 4. The longitudinal associations between number of epigenetic mutations (log10-transformed) and age and other covariates.

Number of epigenetic mutations	Age (year)	Sex (Female to male)	Effect sizes; (p-values)			
			CD19+ B cells (proportion)	1st genetic principal component	Sample quality	Cancer diagnosis
Total epigenetic mutations	8.29e-03 (1.22e-13)	0.0722 (6.33e-03)	4.21 (5.06e-23)	0.445 (0.0413)	0.369 (1.48e-117)	0.0697 (0.0139)
Frequent epigenetic mutations	6.03e-03 (2.17e-19)	-0.0180 (0.33)	1.76 (1.37e-12)	0.595 (1.28e-04)	0.0573 (5.84e-13)	0.0478 (0.0164)
Frequent high methylation outliers	6.81e-03 (2.09e-17)	-0.0314 (0.16)	2.09 (2.25e-12)	0.750 (7.65e-05)	0.0512 (3.58e-08)	0.0602 (0.0130)
Frequent low methylation outliers	2.82e-03 (1.14e-05)	0.0340 (0.057)	0.474 (0.046)	0.0186 (0.92)	0.0888 (8.09e-30)	-6.99e-03 (0.71)

To study epigenetic mutations in relation to cancer, we analyzed DNA methylation data of gene *PRDM7* downloaded from TCGA. Methylation levels of the gene body were much higher in tumor tissues than normal adjacent tissues, but no significant difference in the promoter. Therefore, tumor tissues were observed to have more epigenetic mutations in gene body but not in the promoter. As the normal adjacent tissue is between the normal and tumor tissue, the result suggested that epigenetic mutations may first appear in the promoter in cancer development.

5.3.2 Validation in bisulfite pyrosequencing

As the identified epigenetic mutations could be technical artifacts, we selected one HMO and three LMOs to validate in bisulfite pyrosequencing (described in section 3.3.2). The pyrosequencing results and the corresponding 450k data are presented in Figure 12. In general, methylation levels measure from pyrosequencing were well correlated with that from the 450k array (cg05270750: $r=0.84$; cg17338133: $r=0.59$; cg25351353: $r=0.80$; cg05124918: $r=0.77$). Also, pyrosequencing results showed that epigenetic mutations identified from 450k array were significantly different to normal samples. Therefore, epigenetic mutations identified from the 450k array were successfully validated.

But still, we observed disagreement between pyrosequencing and 450k data in some samples, where four samples in cg17338133 and six samples in cg05124918 showed over 15% difference in methylation level between the two methods. It might indicate that we wrongly detected or failed to detect epigenetic mutations from 450k chip data. In general, changes in methylation levels from pyrosequencing were smoother than that from 450k array (Figure 5). Unlike 450k data, we did not observe the recovery of mutated state back to normal during follow-ups in pyrosequencing. So we suggested that bisulfite pyrosequencing could better present that epigenetic mutations were persistent with and could therefore accumulate in the aging process.

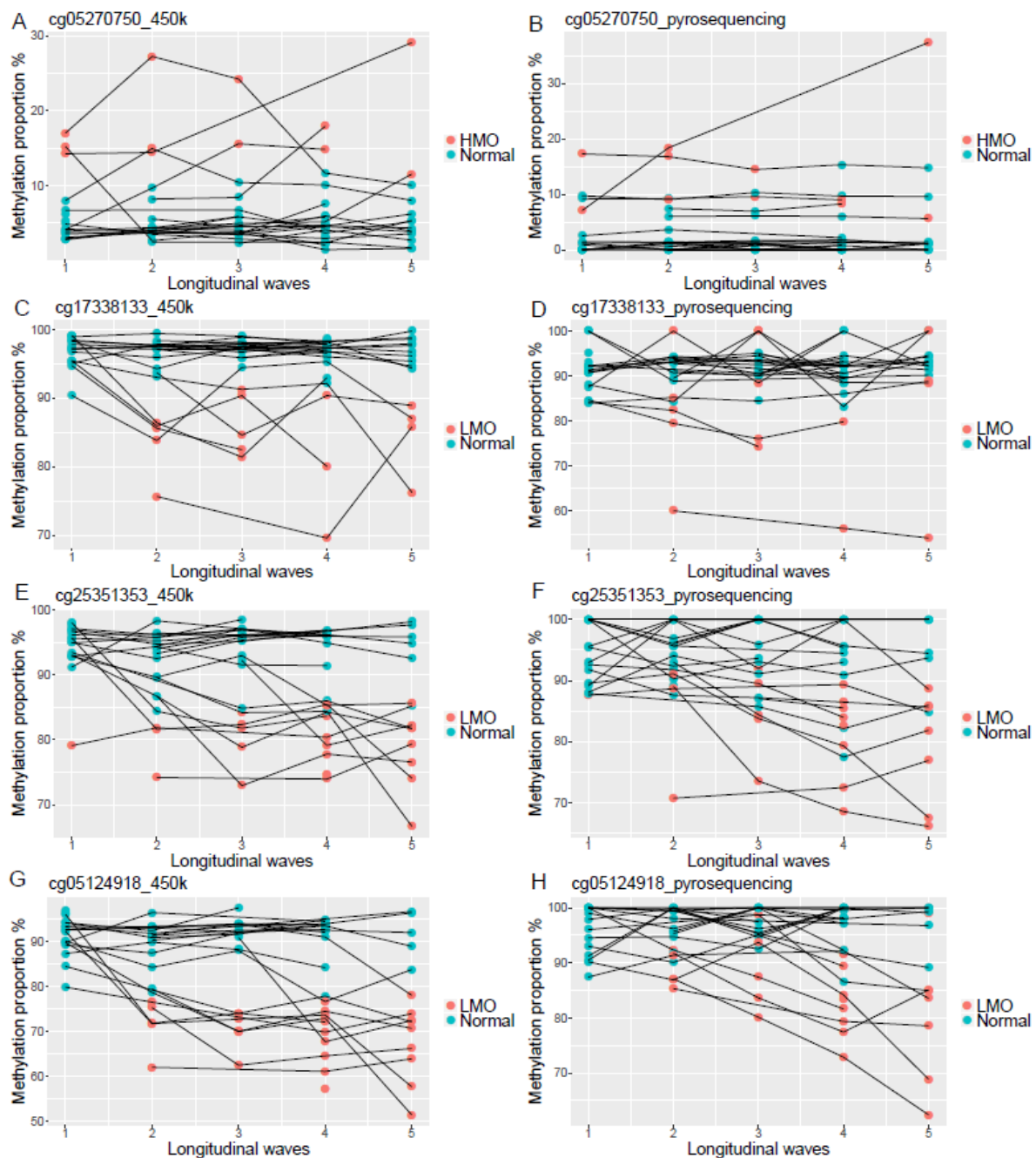


Figure 12. The longitudinal change of four CpGs in 26 individuals measured by 450k array (left panel) and bisulfite pyrosequencing (right panel). Methylation levels of A) cg05270750 from 450k-chip, B) cg05270750 from Pyroseq, C) cg17338133 from 450k-chip, D) cg17338133 from Pyroseq, E) cg25351353 from 450k-chip, F) cg25351353 from Pyroseq, G) cg05124918 from 450k-chip, H) cg05124918 from Pyroseq. Epigenetic mutations identified from 450k array were in red and the reference in blue.

5.4 STUDY IV – DNA METHYLATION AND MORTALITY

5.4.1 EWAS on all-cause and cause-specific mortality

The EWAS on all-cause mortality in SATSA did not identify genome-wide significant association under Bonferroni or FDR correction. Comparing to the three previous publications of EWAS on all-cause mortality, we replicated 9/58 [118], 135/2806 [92] and 42/2552 [93] CpGs with $p < 0.05$ and consistent effect direction respectively.

To explain the inconsistent results between publications, we tested the potential age-varying effect in EWAS. There we identified 6 CpGs with significant age-varying effect on all-cause mortality under Bonferroni correction and 193 CpGs under FDR correction. Of the 193 CpGs, two were identified as age-related CpG in Study I. A number of 13, 231 and 214 CpGs reported from the previous publications [92,93,118] showed age-varying effect under $p < 0.05$. Moreover, we observed the effect of methylation level decreased with age, and eventually changed to the other direction for all CpGs with the age-varying effect. Such change in direction resulted in an average effect close to 0. Two examples of age-varying effect on all-cause mortality are shown in Figure 13.

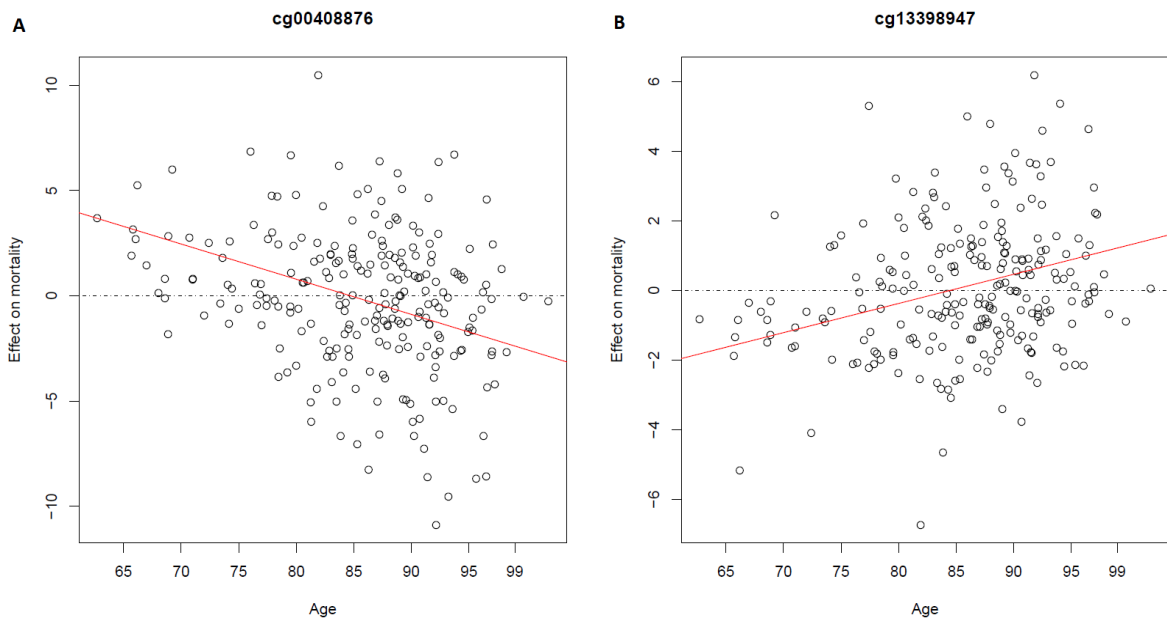


Figure 13. Two examples of age-varying effect of DNA methylation on all-cause mortality. A) cg00408876 ($p = 4.35 \times 10^{-10}$); B) cg13398947 ($p = 7.67 \times 10^{-8}$). Schoenfeld residual plots were used to present the non-proportional hazard ratio over age.

The age-varying effect identified from this study could partly explain the low replication rate between EWAS on all-cause mortality. Without considering the age-varying effect, different studies estimated an average effect of methylation over different ranges of age. This is supported by the results that previously reported CpGs were more found to have age-varying effect than being replicated in the EWAS without age-varying effect. Moreover, the direction of age-varying effect may partly explain that no significant result was identified from the EWAS without age-varying effect.

In the EWAS on cause-specific mortality, we identified 35 CpGs significantly associated with death from cancer under FDR<0.05, but no significant result for death from CVD, stroke or dementia. Genes related to those CpGs showed functions strongly related to cancer, such as genes coding histone methyltransferase, double-strand break repair protein and Insulin like growth factor binding protein. The top 10 CpGs and their gene functions are illustrated in Table 5.

The EWAS results of death from cancer were convincing because their related genes had critical functions in cancer biology. But further evidence is needed to explain their biological mechanisms in relation to cancer mortality. CpGs related to cause-specific mortality may overlap with those related to all-cause mortality, as former outcome is a part of the latter. In this study, cancer-mortality CpGs were related to all-cause mortality to some extent, as 20 of the 35 CpGs showed a p-value lower than 0.05 in the EWAS of all-cause mortality. However, none of them was reported previously [92,93,118].

Table 5. The first 10 significant results from EWAS on death from cancer and their gene functions in relation to cancer.

CpG	HR*	p-value	Gene name	Gene function related to cancer
cg03217966	1.74	1.66e-08	<i>EHMT2</i>	Migratory ability of breast cancer [119]
cg19211619	0.56	4.31e-08	<i>CAPS</i>	RNA expression as prognostic marker in endometrial cancer
cg07380540	0.45	4.87e-08		
cg11136886	0.53	1.79e-07		
cg11919479	0.60	3.42e-07	<i>PAFAH1B3</i>	Critical driver of breast cancer pathogenicity [120]
cg26387956	1.56	4.33e-07	<i>INO80</i>	Oncogenic transcription and tumor growth in non-small cell lung cancer [121]
cg17516160	1.87	4.57e-07	<i>MRE11A</i>	Code double-strand break repair protein
cg19942305	1.69	5.13e-07		
cg24901098	1.77	8.25e-07	<i>IGFBP3</i>	Code insulin like growth factor binding protein
cg06652392	1.79	8.37e-07		

5.4.2 Methylation variability in association with all-cause mortality

Apart from methylation level, we conducted an EWAS to test the methylation variability in associated with all-cause mortality (described in section 4.1.5). We identified 2 significant CpGs under Bonferroni correction and 29 CpGs under FDR correction (Figure 14A). The top CpG is presented in Figure 14B to illustrate the association between methylation variability and mortality risk. The Schoenfeld residual test suggested no violation of proportional hazard assumption among the significant CpGs. Almost all the CpGs with $p < 1 \times 10^{-4}$ had a positive effect size (Figure 14A), which means that people deviate from average methylation levels have a higher mortality risk. One of the 29 CpGs was associated with age as reported in Study I, but none was among the age-varying CpGs reported in Study II.

This study first provided EWAS evidence that methylation variability can be associated with all-cause mortality. Although the positive effect direction was expected, it is not clear how methylation levels of both directions increase mortality. Also, we found age-varying CpGs were not associated with mortality, because the age effect was removed as the time scale in the survival analysis, which reduced the effect of factors strongly correlated with age.

We also tested the number of epigenetic mutations in association with all-cause mortality, but did not observe significant association. Although epigenetic mutations can be important to aging, they did not have strong enough effect on all-cause mortality for us to detect.

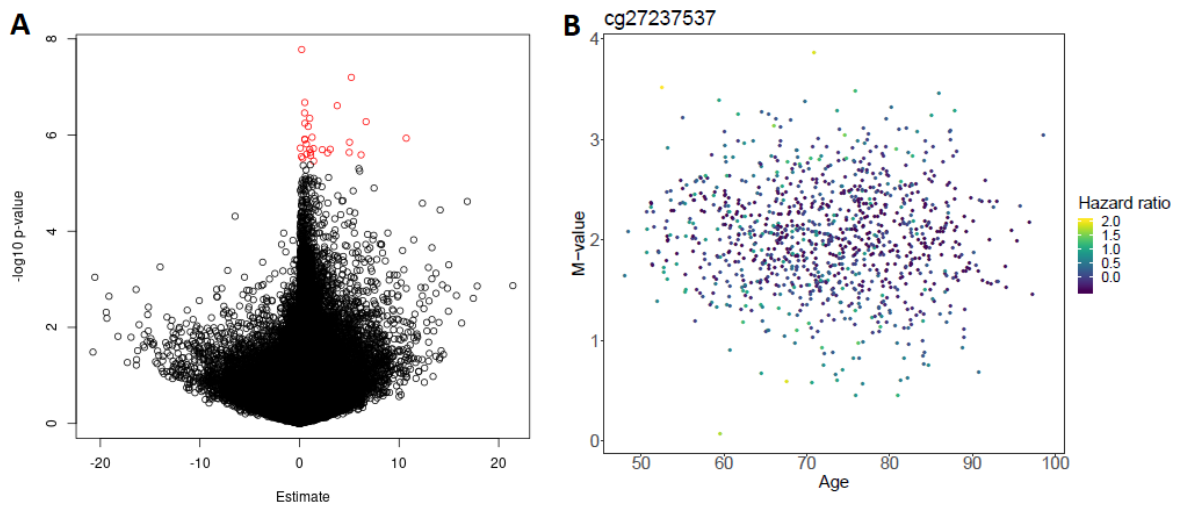


Figure 14. Results from the analysis on methylation variability and all-cause mortality. A) A volcano plot illustrates that most of the CpGs with $p < 1 \times 10^{-4}$ had a positive effect size as colored in red. B) The methylation level of cg27237537, where the variability was associated with all-cause mortality.

6 GENERAL DISCUSSION AND FUTURE PERSPECTIVE

Longitudinal studies can provide a higher level of evidence than cross-sectional studies, and can sometimes imply causation. Repeated measures are particularly important to study DNA methylation as it is dynamic and can be influenced by many factors. Currently, longitudinal studies on DNA methylation are still limited as they are high-cost and time-consuming. The number of repeated measures and the follow-up time are two key factors in longitudinal study, as they make a longitudinal study different to a cross-sectional one. A large number of repeated measurements allowed us to test more hypothesis by modeling complicated changes of DNA methylation over time, such as a quadratic or breakpoint model. Moreover, twins are study subjects matched by nature. By comparing the difference between twins, we can avoid many genetic and environmental confounders. Further studies on discordant twins could help identify factors associated with age-related diseases.

Cellular heterogeneity is one of the major confounder in studies on DNA methylation. Recent evidences suggested that many EWAS signals would be eliminated after properly adjusting for cellular compositions [122]. In most studies, cellular compositions were not measured, but estimated based on methylation data. The quality of methylation data, choice of reference and algorithms can all affect the estimation. Even cell types are properly estimated and adjusted, there are still subtype differences between cells. Future studies on DNA methylation may focus on individual cells, particular for studies on cancer.

Blood samples are the most widely used source of DNA methylation as they are easily available. But some methylation signals may only exist in other tissues and are therefore missing from studies of blood samples. Brain samples are particularly valuable for studies on brain aging and dementia. Also, our study of blood samples identified aging signals related to nervous system. However, since brain samples are mostly donated from deceased people, DNA started to breakdown after death making it hard to measure DNA methylation.

Similarly, genetic has a strong effect on DNA methylation and DNA methylation may mediate the genetic effect on traits. However, the microarray technique cannot properly measure CpGs that are too close to a SNP, or the CpGs themselves are SNPs. Most studies simply removed these CpGs as a standard QC procedure. Moreover, rare genetic variants may have stronger effect on DNA methylation but are often ignored. Therefore, the genetic effect on DNA methylation is underestimated. Future studies using whole-genome bisulfite sequencing can better measure the genetic effect on DNA methylation.

DNA methylation can be affected by so many factors that we probability cannot properly adjust all confounders. For GWAS, the Q-Q plot is a powerful method to diagnose unadjusted confounders by testing how p-values deviate from a uniform distribution. However, the tests in EWAS are not independent so that we expect to see an early deviation in the QQ plot when modeling strong factors like age. Further studies are needed to have a better understanding of the correlation structure between CpGs and provide a theoretical distribution of test results given a hypothesized association.

Many CpGs are highly correlated to their neighbors and the alteration of methylation can happen at multiple adjacent CpG sites [123]. But still, there is no consensus on local CpG regions. In addition, the long-range interaction between CpGs makes it more complex to cluster CpGs into groups. The long-range interaction, sometime between chromosomes, is a result of physically contact in nucleosome. Thus, studies of chromosome conformation capture can help establish the methylation pattern of long-range interactions.

The regulatory and functional annotations of DNA methylation are essential to understand the properties of CpGs identified from EWAS. However, the target genes of regulatory regions are still unclear, because CpGs may not regulate the nearest gene, but regulate one or more distal genes. This is supported by the fact that CpGs may not correlated with the expression of genes close by. Therefore, the better approach is to integrate DNA methylation with other epigenetic markers and omics data, in order to determine regulatory functions and study their interplays.

7 CONCLUSION

In this thesis, we used repeated measures from the SATSA study to investigate mechanisms of DNA methylation related to aging from a longitudinal perspective. The mechanisms included changes in average methylation level, increase in inter-individual methylation variability and increase in the number of epigenetic mutations with age. Further, we analyzed these mechanisms in association with all-cause and cause-specific mortality.

In Study I, we identified age-associated CpGs from a longitudinal EWAS and validated the results in external cohorts. The analysis on twins and meQTLs found genetic effect on age-associated CpGs, but age and genetic effects were independent.

In Study II, we developed a method to model the longitudinal change of methylation variability with age. We then applied the method an EWAS and identified age-varying CpGs from longitudinal data.

In Study III, we verified that the number of epigenetic mutations increased with age in longitudinal data, and identified other associated factors. We further validated results using bisulfite pyrosequencing and proved that epigenetic mutations were persistent and could accumulate with age.

In Study IV, we observed the age-varying effect of DNA methylation on all-cause mortality, which might explain inconsistent results from previous publications. We also identified CpGs associated with death from cancer and their gene functions were critical to cancer. Furthermore, we provided longitudinal evidence that methylation variability could also predict all-cause mortality.

8 ACKNOWLEDGEMENT

The four-year PhD time at the MEB department was a wonderful experience for me. Here I would like to thank my supervisors, colleagues, collaborators and friends for your help and good company.

First and foremost, I would like to thank **Sara Hägg**, my main supervisor, for bringing me in your group and guiding me in scientific research. I am always encouraged by you to broaden my knowledge and perform independent research. Also I am appreciate that you provided me great opportunities to communicate and collaborate with other researchers. You are also a great leader and more importantly a nice person. It is truly my pleasure to study from you and work with you.

Robert Kalsson, my co-supervisor and statistician, thank you for always being available to answer my questions. Your deep knowledge in genetics and excellent programming skills inspired me to develop my skill set.

Åsa Hedman, my co-supervisor, thank you for your support in DNA methylation. Your rich experience and the brilliant ideas you brought up greatly improved my studies. Especially, I want to thank you for always being patient to comment on my manuscripts.

Malin Almgren, my co-supervisor, thank you for your teaching and support on techniques of biotechnology. I am so grateful for you patient help on my bisulfite pyrosequencing. It was such a pleasure working with you in lab.

Catarina Almqvist, my co-supervisor, thank you for your suggestions on my study design. You have always been open to provide me helps. I really appreciate your support and encouragement.

It is a great pleasure for me to work in the fantastic Pedersen's Group. Frist, I would like to thank our group leader **Nancy Pedersen**. I am always inspired by your broad and solid scientific knowledge, and charming personality. It is my honor to work on the projects you started. And thank you for the scientific and language support in my research work. Also, I wish to thank our awesome group members, **Fei Yang, Kathleen Bokenberger, Catalina Zavala, Xu Chen, Bojing Liu, Kelli Lehto, Dylan Williams, Miriam Mosing, Robert Miller, Xia Li, Johanna Sieurin, KK, Ge Bai, Qi Wang, Yasutake Tomata, Karolina Kauppi, Kristina Johnell and Mat é Szilcz**. We could not have such a great working environment without all of your contributions. Especially, I would like to thank **Juulia Jylh ä v ä**, for our collaborations in methylation studies and epigenetic clock. **Yiqiang Zhan**, for our nice chatting when we were sharing a room. **Malin Ericsson**, for all your kindly help in my whole PhD time. **Ida Karlsson**, for your patient help in registry data and involving me in you study.

I would also like to thank collaborators in my research. **Chandra Reynolds** for involving me your research and our nice discussions. **Patrik Magnusson**, for your enthusiasm and humor.

And to the lovely MEB **Biostatistics group**, thank you all for being friendly and helpful. Your research sparked my interest in statistics and inspired me to learn more. And, it was truly wonderful to have so many great seminars, fika and special retreats.

Also, to all my **Chinese friends** at MEB. Thank you all creating such a warm community that is always helpful.

Moreover, I want to thank our TA stuffs in MEB for supporting our research. You have help make MEB such a great place to work. Especially, I would like to thank **Camilla Ahlqvist**, for your support and collaboration when I was the student seminar coordinator. **Rikard Öberg**, for your help in the Unix system and your work to build and maintain the MEB sever. **Erika Nordenhagen** and **Janina Mahmoodi**, for your help on our group.

In the end, I would love to thank my **family** in China for always supporting me and encouraging me to keep going. And to **Yinxi**, my love, thank you for coming to my life. We will enjoy our life journey together.

9 REFERENCE

1. Dupont C, Armant DR, Brenner CA. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Semin Reprod Med.* 2009 Sep;27(5):351–7.
2. Waddington CH. Genetic Assimilation of the Bithorax Phenotype. *Evolution.* 1956;10(1):1–13.
3. Felsenfeld G. A Brief History of Epigenetics. *Cold Spring Harb Perspect Biol.* 2014 Jan 1;6(1):a018200.
4. Brockdorff N, Turner BM. Dosage Compensation in Mammals. *Cold Spring Harb Perspect Biol.* 2015 Mar 1;7(3):a019406.
5. Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell.* 2004 Jan 23;116(2):281–97.
6. Kumar S, Chinnusamy V, Mohapatra T. Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Front Genet.* 2018;9:640.
7. Fatemi M, Pao MM, Jeong S, Gal-Yam EN, Egger G, Weisenberger DJ, et al. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic Acids Res.* 2005;33(20):e176.
8. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A.* 2006 Jan 31;103(5):1412–7.
9. Deaton AM, Webb S, Kerr ARW, Illingworth RS, Guy J, Andrews R, et al. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res.* 2011 Jul;21(7):1074–86.
10. Robertson KD, Keyomarsi K, Gonzales FA, Velicescu M, Jones PA. Differential mRNA expression of the human DNA methyltransferases (DNMTs) 1, 3a and 3b during the G0/G1 to S phase transition in normal and tumor cells. *Nucleic Acids Res.* 2000 May 15;28(10):2108–13.
11. Okano M, Xie S, Li E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genetics.* 1998 Jul 1;19(3):ng0798_219.
12. Okano M, Bell DW, Haber DA, Li E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell.* 1999 Oct 29;99(3):247–57.
13. Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev.* 2014 Apr 15;28(8):812–28.
14. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature.* 2013 Oct 24;502(7472):472–9.
15. Nan X, Cross S, Bird A. Gene silencing by methyl-CpG-binding proteins. *Novartis Found Symp.* 1998;214:6–16; discussion 16–21, 46–50.
16. Angeloni A, Bogdanovic O. Enhancer DNA methylation: implications for gene regulation. *Essays Biochem.* 2019 Sep 24;

17. Clermont P-L, Parolia A, Liu HH, Helgason CD. DNA methylation at enhancer regions: Novel avenues for epigenetic biomarker development. *Front Biosci (Landmark Ed)*. 2016 Jan 1;21:430–46.
18. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. On the presence and role of human gene-body DNA methylation. *Oncotarget*. 2012 May 9;3(4):462–74.
19. Bogdanović O, Lister R. DNA methylation and the preservation of cell identity. *Curr Opin Genet Dev*. 2017 Oct;46:9–14.
20. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*. 2014 Feb 4;15(2):R31.
21. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012 May 8;13(1):86.
22. Zheng SC, Beck S, Jaffe AE, Koestler DC, Hansen KD, Houseman AE, et al. Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. *Nat Meth*. 2017 Mar;14(3):216–7.
23. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*. 2011;12:R10.
24. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet*. 2011 Apr 8;88(4):450–7.
25. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *PNAS*. 2005 Jul 26;102(30):10604–9.
26. Relton CL, Davey Smith G. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med*. 2010 Oct 26;7(10):e1000356.
27. Grundberg E, Meduri E, Sandling JK, Hedman ÅK, Keildson S, Buil A, et al. Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements. *The American Journal of Human Genetics*. 2013 Nov 7;93(5):876–90.
28. Bell JT, Tsai P-C, Yang T-P, Pidsley R, Nisbet J, Glass D, et al. Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population. *PLoS Genet*. 2012 Apr 19;8(4):e1002629.
29. McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biology*. 2014;15:R73.
30. Gordon L, Joo JE, Powell JE, Ollikainen M, Novakovic B, Li X, et al. Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Res*. 2012 Aug;22(8):1395–406.
31. van Dongen J, Nivard MG, Willemsen G, Hottenga J-J, Helmer Q, Dolan CV, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications*. 2016 Apr 7;7:11115.

32. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*. 2016;17:61.
33. Lemire M, Zaidi SHE, Ban M, Ge B, A ĩsi D, Germain M, et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nature Communications*. 2015 Feb 26;6:ncomms7326.
34. Fan J-B, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, et al. Highly Parallel SNP Genotyping. *Cold Spring Harb Symp Quant Biol*. 2003 Jan 1;68:69–78.
35. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. Whole-genome genotyping with the single-base extension assay. *Nature Methods*. 2006 Jan;3(1):31–3.
36. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010 Nov 30;11(1):587.
37. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011 Oct;98(4):288–95.
38. O’Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016 Jan 4;44(D1):D733-745.
39. Forest M, O’Donnell KJ, Voisin G, Gaudreau H, MacIsaac JL, McEwen LM, et al. Agreement in DNA methylation levels from the Illumina 450K array across batches, tissues, and time. *Epigenetics*. 2018 Jan 30;13(1):19–32.
40. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-Wide Association Studies for common human diseases. *Nat Rev Genet*. 2011 Jul 12;12(8):529–41.
41. Pe’er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*. 2008;32(4):381–5.
42. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol*. 2013 Mar 12;14(3):R21.
43. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *PNAS*. 2003 Aug 5;100(16):9440–5.
44. Johnson FB, Sinclair DA, Guarente L. Molecular biology of aging. *Cell*. 1999 Jan 22;96(2):291–302.
45. López-Ot ĩn C, Blasco MA, Partridge L, Serrano M, Kroemer G. The Hallmarks of Aging. *Cell*. 2013 Jun 6;153(6):1194–217.
46. Sidler C, Kovalchuk O, Kovalchuk I. Epigenetic Regulation of Cellular Senescence and Aging. *Front Genet*. 2017;8:138.
47. von Zglinicki T, Martin-Ruiz CM. Telomeres as biomarkers for ageing and age-related diseases. *Curr Mol Med*. 2005 Mar;5(2):197–203.
48. Gorbunova V, Seluanov A, Mao Z, Hine C. Changes in DNA repair during aging. *Nucleic Acids Res*. 2007;35(22):7466–74.

49. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, et al. Distinct DNA methylomes of newborns and centenarians. *PNAS*. 2012 Jun 26;109(26):10522–7.
50. Powers ET, Morimoto RI, Dillin A, Kelly JW, Balch WE. Biological and chemical approaches to diseases of proteostasis deficiency. *Annu Rev Biochem*. 2009;78:959–91.
51. Nogueiras R, Habegger KM, Chaudhary N, Finan B, Banks AS, Dietrich MO, et al. Sirtuin 1 and sirtuin 3: physiological modulators of metabolism. *Physiol Rev*. 2012 Jul;92(3):1479–514.
52. Kujoth GC, Hiona A, Pugh TD, Someya S, Panzer K, Wohlgemuth SE, et al. Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science*. 2005 Jul 15;309(5733):481–4.
53. Campisi J, d’Adda di Fagagna F. Cellular senescence: when bad things happen to good cells. *Nat Rev Mol Cell Biol*. 2007 Sep;8(9):729–40.
54. Blagosklonny MV. Aging: ROS or TOR. *Cell Cycle*. 2008 Nov 1;7(21):3344–54.
55. Vaziri H, Benchimol S. From telomere loss to p53 induction and activation of a DNA-damage pathway at senescence: the telomere loss/DNA damage model of cell aging. *Exp Gerontol*. 1996 Apr;31(1–2):295–301.
56. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell*. 2011 Mar 4;144(5):646–74.
57. Gentilini D, Garagnani P, Pisoni S, Bacalini MG, Calzari L, Mari D, et al. Stochastic epigenetic mutations (DNA methylation) increase exponentially in human aging and correlate with X chromosome inactivation skewing in females. *Aging (Albany NY)*. 2015 Aug;7(8):568–78.
58. Zampieri M, Ciccarone F, Calabrese R, Franceschi C, Bürkle A, Caiafa P. Reconfiguration of DNA methylation in aging. *Mech Ageing Dev*. 2015 Nov;151:60–70.
59. Huidobro C, Fernandez AF, Fraga MF. Aging epigenetics: causes and consequences. *Mol Aspects Med*. 2013 Aug;34(4):765–81.
60. Richardson B. Impact of aging on DNA methylation. *Ageing Research Reviews*. 2003 Jul;2(3):245–61.
61. Rakyan VK, Down TA, Maslau S, Andrew T, Yang T-P, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res*. 2010 Apr;20(4):434–9.
62. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, et al. Distinct DNA methylomes of newborns and centenarians. *PNAS*. 2012 Jun 26;109(26):10522–7.
63. McClay JL, Aberg KA, Clark SL, Nerella S, Kumar G, Xie LY, et al. A methylome-wide study of aging using massively parallel sequencing of the methyl-CpG-enriched genomic fraction from blood in over 700 subjects. *Hum Mol Genet*. 2014 Mar 1;23(5):1175–85.
64. Sheaffer KL, Elliott EN, Kaestner KH. DNA hypomethylation contributes to genomic instability and intestinal cancer initiation. *Cancer Prev Res (Phila)*. 2016 Jul;9(7):534–46.
65. Peters I, Vaske B, Albrecht K, Kuczyk MA, Jonas U, Serth J. Adiposity and age are statistically related to enhanced RASSF1A tumor suppressor gene promoter methylation in normal autopsy kidney tissue. *Cancer Epidemiol Biomarkers Prev*. 2007 Dec;16(12):2526–32.

66. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, et al. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*. 2013 Jan 24;49(2):359–67.
67. Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, et al. Age-associated DNA methylation in pediatric populations. *Genome Res*. 2012 Apr;22(4):623–32.
68. Johansson A, Enroth S, Gyllenstein U. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLoS ONE*. 2013;8(6):e67378.
69. Florath I, Butterbach K, Muller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Human Molecular Genetics*. 2014 Mar 1;23(5):1186–201.
70. Garagnani P, Bacalini MG, Pirazzini C, Gori D, Giuliani C, Mari D, et al. Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*. 2012 Dec 1;11(6):1132–4.
71. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics*. 2018 Jun;19(6):371–84.
72. Jylhävä J, Pedersen NL, Hägg S. Biological Age Predictors. *EBioMedicine*. 2017 Apr 1;21:29–36.
73. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biology*. 2013 Oct 21;14(10):R115.
74. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, et al. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology*. 2015;16:25.
75. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai P-C, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY)*. 2016 Sep 28;8(9):1844–65.
76. Perna L, Zhang Y, Mons U, Holleczer B, Saum K-U, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clinical Epigenetics*. 2016 Jun 3;8:64.
77. Marioni RE, Shah S, McRae AF, Ritchie SJ, Muniz-Terrera G, Harris SE, et al. The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *Int J Epidemiol*. 2015 Aug;44(4):1388–96.
78. Breitling LP, Saum K-U, Perna L, Schöttker B, Holleczer B, Brenner H. Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. *Clin Epigenetics*. 2016;8:21.
79. Quach A, Levine ME, Tanaka T, Lu AT, Chen BH, Ferrucci L, et al. Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging (Albany NY)*. 2017 Feb 14;9(2):419–46.
80. Levine ME, Lu AT, Bennett DA, Horvath S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging (Albany NY)*. 2015 Dec;7(12):1198–211.
81. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. *Aging (Albany NY)*. 2015 Dec;7(12):1130–42.

82. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)*. 2018 18;10(4):573–91.
83. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging (Albany NY)*. 2018 Jul 26;10(7):1758–75.
84. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)*. 2019 21;11(2):303–27.
85. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *PNAS*. 2005 Jul 26;102(30):10604–9.
86. Wang Y, Karlsson R, Lampa E, Zhang Q, Hedman ÅK, Almgren M, et al. Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins. *bioRxiv*. 2017 Nov 29;226266.
87. Slieker RC, van Iterson M, Luijk R, Beekman M, Zhernakova DV, Moed MH, et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol*. 2016 Sep 22;17(1):191.
88. Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, et al. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nature Communications*. 2016 Jan 29;7:10478.
89. Teschendorff AE, Jones A, Widschwendter M. Stochastic epigenetic outliers can define field defects in cancer. *BMC Bioinformatics*. 2016 Apr 22;17(1):178.
90. Feinberg AP, Koldobskiy MA, Gündör A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat Rev Genet*. 2016 May;17(5):284–99.
91. Gentilini D, Scala S, Gaudenzi G, Garagnani P, Capri M, Cescon M, et al. Epigenome-wide association study in hepatocellular carcinoma: Identification of stochastic epigenetic mutations through an innovative statistical approach. *Oncotarget*. 2017 Jun 27;8(26):41890–902.
92. Svane AM, Soerensen M, Lund J, Tan Q, Jylhävä J, Wang Y, et al. DNA Methylation and All-Cause Mortality in Middle-Aged and Elderly Danish Twins. *Genes (Basel)*. 2018 Feb 8;9(2).
93. Lund JB, Li S, Baumbach J, Svane AM, Hjelmberg J, Christiansen L, et al. DNA methylome profiling of all-cause mortality in comparison with age-associated methylation patterns. *Clinical Epigenetics*. 2019 Feb 8;11(1):23.
94. Finkel D, Pedersen NL. Processing Speed and Longitudinal Trajectories of Change for Cognitive Abilities: The Swedish Adoption/Twin Study of Aging. *Aging, Neuropsychology, and Cognition*. 2004 Jun 1;11(2–3):325–45.
95. Lind L, Fors N, Hall J, Marttala K, Stenborg A. A Comparison of Three Different Methods to Evaluate Endothelium-Dependent Vasodilation in the Elderly. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2005 Nov 1;25(11):2368–75.
96. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936. *Int J Epidemiol*. 2012 Dec 1;41(6):1576–84.
97. Morris TJ, Beck S. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods*. 2015 Jan 15;72:3–8.

98. Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Meth.* 2014 Nov;11(11):1138–40.
99. Yousefi P, Huen K, Aguilar Schall R, Decker A, Elboudwarej E, Quach H, et al. Considerations for Normalization of DNA Methylation Data by Illumina 450K BeadChip Assay in Population Studies. *Epigenetics.* 2013 Nov;8(11):1141–52.
100. Wang T, Guan W, Lin J, Boutaoui N, Canino G, Luo J, et al. A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. *Epigenetics.* 2015 Jun 2;10(7):662–9.
101. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* 2013 Apr;41(7):e90.
102. Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics.* 2013 May 1;14(1):293.
103. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, et al. Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS ONE.* 2012 Jul 25;7(7):e41361.
104. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007 Jan;8(1):118–27.
105. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012 Mar 15;28(6):882–3.
106. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics.* 2009 Jun 19;5(6):e1000529.
107. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature.* 2015 Oct 1;526(7571):68–74.
108. D éz-Villanueva A, Mallona I, Peinado MA. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics & Chromatin.* 2015 Jun 23;8(1):22.
109. Breusch TS, Pagan AR. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica.* 1979;47(5):1287–94.
110. Shabalín AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics.* 2012 May 15;28(10):1353–8.
111. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The Ensembl Regulatory Build. *Genome Biology.* 2015 Mar 24;16(1):56.
112. Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature.* 2011 May;473(7345):43–9.
113. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4(1):44–57.

114. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009 Jan;37(1):1–13.
115. Sturm G, Cardenas A, Bind M-A, Horvath S, Wang S, Wang Y, et al. Human aging DNA methylation signatures are conserved but accelerated in cultured fibroblasts. *Epigenetics.* 2019 Jun 3;0(0):1–16.
116. Tan Q, Heijmans BT, Hjelmborg J v B, Soerensen M, Christensen K, Christiansen L. Epigenetic drift in the aging genome: a ten-year follow-up in an elderly twin cohort. *Int J Epidemiol.* 2016 Aug 1;45(4):1146–58.
117. Ghosh J, Schultz B, Coutifaris C, Sapienza C. Highly variant DNA methylation in normal tissues identifies a distinct subclass of cancer patients. *Adv Cancer Res.* 2019;142:1–22.
118. Zhang Y, Wilson R, Heiss J, Breitling LP, Saum K-U, Schöttker B, et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nature Communications.* 2017 Mar 17;8:14617.
119. Kim K, Son M-Y, Jung C-R, Kim D-S, Cho H-S. EHMT2 is a metastasis regulator in breast cancer. *Biochemical and Biophysical Research Communications.* 2018 Feb 5;496(2):758–62.
120. Mulvihill MM, Benjamin DI, Ji X, Le Scolan E, Louie SM, Shieh A, et al. Metabolic Profiling Reveals PAFAH1B3 as a Critical Driver of Breast Cancer Pathogenicity. *Chem Biol.* 2014 Jul 17;21(7):831–40.
121. Zhang S, Zhou B, Wang L, Li P, Bennett BD, Snyder R, et al. INO80 is required for oncogenic transcription and tumor growth in non-small cell lung cancer. *Oncogene.* 2017;36(10):1430–9.
122. Kong Y, Rastogi D, Seoighe C, Grealley JM, Suzuki M. Insights from deconvolution of cell subtype proportions enhance the interpretation of functional genomic data. *PLOS ONE.* 2019 Apr 25;14(4):e0215987.
123. Lökvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Res.* 2016 20;44(11):5123–32.

